

---

# **GENE DUPLICATION**

---

Edited by **Felix Friedberg**

**INTECHWEB.ORG**

## **Gene Duplication**

Edited by Felix Friedberg

## **Published by InTech**

Janeza Trdine 9, 51000 Rijeka, Croatia

## **Copyright © 2011 InTech**

All chapters are Open Access articles distributed under the Creative Commons Non Commercial Share Alike Attribution 3.0 license, which permits to copy, distribute, transmit, and adapt the work in any medium, so long as the original work is properly cited. After this work has been published by InTech, authors have the right to republish it, in whole or part, in any publication of which they are the author, and to make other personal use of the work. Any republication, referencing or personal use of the work must explicitly identify the original source.

Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published articles. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

**Publishing Process Manager** Martina Blečić

**Technical Editor** Teodora Smiljanic

**Cover Designer** Jan Hyrat

**Image Copyright** Booka, 2011. Used under license from Shutterstock.com

First published September, 2011

Printed in Croatia

A free online edition of this book is available at [www.intechopen.com](http://www.intechopen.com)  
Additional hard copies can be obtained from [orders@intechweb.org](mailto:orders@intechweb.org)

Gene Duplication, Edited by Felix Friedberg

p. cm.

ISBN 978-953-307-387-3

**INTECH** OPEN ACCESS  
PUBLISHER

**INTECH** open

**free** online editions of InTech  
Books and Journals can be found at  
**[www.intechopen.com](http://www.intechopen.com)**





---

# Contents

---

## **Preface IX**

### **Part 1 General Aspects 1**

- Chapter 1 **A Theoretical Scheme of the Large-Scale Evolution by Generating New Genes from Gene Duplication 3**  
Jinya Otsuka
- Chapter 2 **Duplicated Gene Evolution Following Whole-Genome Duplication in Teleost Fish 27**  
Baocheng Guo, Andreas Wagner and Shunping He
- Chapter 3 **Detection and Analysis of Functional Specialization in Duplicated Genes 37**  
Owen Z. Woody and Brendan J. McConkey
- Chapter 4 **Predicting Tandemly Arrayed Gene Duplicates with WebScipio 59**  
Klas Hatje and Martin Kollmar
- Chapter 5 **The LRR and TM Containing Multi-Domain Proteins in Arabidopsis 77**  
Felix Friedberg
- Chapter 6 **Partial Gene Duplication and the Formation of Novel Genes 95**  
Macarena Toll-Riera, Steve Laurie, Núria Radó-Trilla and M.Mar Albà

### **Part 2 A Look at Some Gene Families 111**

- Chapter 7 **Immunoglobulin Polygeny: An Evolutionary Perspective 113**  
J. E. Butler, Xiu-Zhu Sun and Nancy Wertz

Chapter 8	<b>Gene Duplication in Insecticide Resistance</b> 141
	Si Hyeock Lee and Deok Ho Kwon
Chapter 9	<b>Gene Duplication and the Origin of Translation Factors</b> 151
	Galina Zhouravleva and Stanislav Bondarev
Chapter 10	<b>Analysis of Duplicate Gene Families in Microbial Genomes and Application to the Study of Gene Duplication in <i>M. tuberculosis</i></b> 173
	Venu Vuppu and Nicola Mulder
Chapter 11	<b>The Evolutionary History of CBF Transcription Factors: Gene Duplication of CCAAT – Binding Factors NF-Y in Plants</b> 197
	Alexandro Cagliari, Andreia Carina Turchetto-Zolet, Felipe dos Santos Maraschin, Guilherme Loss, Rogério Margis and Marcia Margis-Pinheiro
<b>Part 3</b>	<b>Examining Bundles of Genes</b> 223
Chapter 12	<b>L- Myo-Inositol 1-Phosphate Synthase (MIPS) in Chickpea: Gene Duplication and Functional Divergence</b> 225
	Manoj Majee and Harmeet Kaur
Chapter 13	<b>On the Specialization History of the ADP-Dependent Sugar Kinase Family</b> 237
	Felipe Merino and Victoria Guixé
Chapter 14	<b>Duplication of Coagulation Factor Genes and Evolution of Snake Venom Prothrombin Activators</b> 257
	Shiyang Kwong and R. Manjunatha Kini
Chapter 15	<b>A Puroindoline Mutigene Family Exhibits Sequence Diversity in Wheat and is Associated with Yield-Related Traits</b> 279
	Feng Chen, Fuyan Zhang, Craig F. Morris and Dangqun Cui
Chapter 16	<b>Evolution of GPI-Aspartyl Proteinases (Yapsines) of <i>Candida</i> spp</b> 289
	Berenice Parra-Ortega, Lourdes Villa-Tanaca and César Hernández-Rodríguez
Chapter 17	<b>Clues to Evolution of the SERA Multigene Family in the Genus <i>Plasmodium</i></b> 315
	Nobuko Arisue, Nirianne M. Q. Palacpac, Kazuyuki Tanabe and Toshihiro Horii

- Chapter 18 **Molecular Evolution of Juvenile Hormone Signaling** 333  
Aaron A. Baumann and Thomas G. Wilson
- Chapter 19 **Gene Duplication and Subsequent Differentiation  
of Esterases in Cactophilic *Drosophila* Species** 353  
Rogério P. Mateus, Luciana P. B. Machado and Carlos R. Ceron
- Chapter 20 ***SNCA* Gene Multiplication:  
A Model Mechanism of Parkinson Disease** 373  
Kenya Nishioka, Owen A. Ross and Nobutaka Hattori
- Chapter 21 **Bucentaur (*Bcnt*) Gene Family:  
Gene Duplication and Retrotransposon Insertion** 383  
Shintaro Iwashita and Naoki Osada



---

## Preface

---

The book *Gene Duplication* consists of 21 chapters divided in 3 parts: General Aspects, A Look at Some Gene Families and Examining Bundles of Genes.

The importance of the study of Gene Duplication stems from the realization that the dynamic process of duplication is the “*sine qua non*” underlying the evolution of all living matter. Genes may be altered before or after the duplication process thereby undergoing neofunctionalization, thus creating in time new organisms which populate the Earth.

Osaka (Chapter 1) suggests that similarities in amino acid sequences exhibited by paralogous proteins prove that evolution proceeds via *in toto* gene duplication. If the ancestral and the newly created gene perform the same function, the new gene would be labeled a subfunctional gene. It should be added that such a duplicated gene encoding an identical product might also be engaged by different cellular regulatory signals (e.g. methylation of nucleotide sites) which in turn, could hamper the expression of such a duplicated gene. (See e.g. Woody et al. Chapter 3). If this duplicated gene subsequently undergoes mutations that allow a function for the new gene that is different from the parent gene (neofunctionalization) that would represent a far more positive evolutionary event. The first three chapters in this book focus on such *in toto* gene duplications whereby in evolutionary time neofunctionalization could have taken hold. There are also several specific circumscribed examples given in this book. (See e.g. Majee&Kaur, Chapter 12). Undoubtedly, duplication contributes substantially to the formation of new genes. But there is a caveat: In time, the majority of duplicated genes mutates into oblivion.

In recent years, however, attention has been paid to another possible path for creating a new gene: The formation of the chimeric gene, a gene immediately ready for a new function. Such a gene might result from altering the position of spliced introns, or more likely from retropositioning of a new encoding domain into the gene: I.e. partial gene duplications and combination. It is obvious that such processes are particularly suited for the creation of genes encoding multi-domain proteins and that they may accelerate considerably the natural process of neofunctionalization. (See Hatje et al. Chapter 4; Friedberg, Chapter 5; Toll-Riera et al. Chapter 6 and Iwashita et al. Chapter 21). Retrotransposons are capable of promoting such segmental duplications.

"Retroduplication" contributes significantly to the formation of new genes. These genes, in turn may also be duplicated and eventually be erased into oblivion by mutations.

**Prof. Felix Friedberg**

Howard University Medical School,  
Washington DC,  
USA







# **Part 1**

## **General Aspects**



# A Theoretical Scheme of the Large-Scale Evolution by Generating New Genes from Gene Duplication

Jinya Otsuka  
*JO Institute of Biophysics, Tokyo,  
Japan*

## 1. Introduction

In the famous book “The Origin of Species” by Darwin (1859), the gradual accumulation of selectively advantageous variants has been proposed qualitatively by obtaining a hint from the artificial selection of domestic animals and plants as well as from the observation of unique species in a geographically isolated region. The core of this proposal has become evident, after the re-discovery of Mendelian heredity, by the detection of hereditary variants, i. e., mutants, and extensive investigations have been carried out for the behavior of mutants especially in the *Drosophila* population (Dobzhansky, 1941; Mayer, 1942; Huxley, 1943; Simpson, 1944). In parallel, Darwinian evolution is mathematically formulated in population genetics to estimate the probability that a spontaneously generated mutant is fixed in, or eliminated from, the population according to the positive or negative value of a selective parameter (Fisher, 1930; Wright, 1949). Although the accumulation of such mutants as those found in the *Drosophila* was supposed to explain the whole process of evolution, the mutants detected at that time were mainly due to the point mutations in established genes, and most of them were defective. Thus, doubts remain about whether the gradual accumulation of such mutants gives rise to radically new organs such as wings and eyes. Another criticism against the survival of the fittest in Darwinian evolution is also raised by the ecological fact of diversity that different styles of organisms coexist in the same area (Nowak et al., 1994).

The gene and genome sequencing, which started in the latter half of the last century, has brought new information about the evolution of organisms. First, the amino acid sequence similarities of paralogous proteins strongly suggest that the repertoire of protein functions has been expanded by gene duplication, succeeding nucleotide base substitutions, partial insertion and deletion, and further by domain shuffling in some cases (Ingram, 1963; Gilbert, 1978; Ferris & White, 1979). Such examples are now increasing, proposing many protein families and superfamilies. Second, the clustering analysis of proteomes reveals a characteristic feature that the proteins functioning in the core part are essentially common to both prokaryotes and eukaryotes, and that the decisive difference in gene repertoire between the organisms is observed in the peripheral parts displaying different living styles (Kojima & Otsuka, 2000 a, b, c; Kojima & Otsuka, 2002). These sequence data are now compiled into databases (e. g., Wheeler et al., 2004; Birney et al., 2006).

Although the importance of gene duplication in evolution was already indicated in the last century (Ohno, 1970), this indication still remained describing the circumstantial evidence of gene duplication and the fossil record of vertebrate organs in a qualitative way. Theoretically, some new concept is needed to formulate the evolution by gene duplication, going beyond the narrow view of population genetics which only focuses on a mutated gene. For this purpose, the author has recently proposed the new concept of 'biological activity', which is determined by a whole genome, and explained the divergence of the original style of organisms and the new style of organisms having a new gene generated from the counterpart of duplicated genes (Otsuka, 2005; 2008). This evolution by gene duplication will be called the large-scale evolution, being distinguished from Darwinian evolution.

In this chapter, the explanatory remarks are first given for the concept of 'biological activity' and the large-scale of evolution will be then investigated in detail on the three types of organisms, which are different in their genome constitution and transmission. The genome is a single DNA molecule in most prokaryotes and it is a set of chromosomes in lower eukaryotes. These organisms will be tentatively called the monoploid organisms as the first type of organisms. Some lower eukaryotes exchange homologous chromosomes through the process of conjugation. These lower eukaryotes are treated as the second type of organisms. In higher animals and plants, each of the cells constituting the adult form carries the genome consisting of the plural number of homologous chromosome pairs, and the monoploid state only appears in the gametes (egg and sperm). These higher eukaryotes will be treated as the third type, being called the diploid organisms in the sense that the present study focuses on the evolution of the characters expressed in their diploid state. The main purpose of the present study is to elucidate the difference between the three types of organisms, especially in the probabilities that two or more kinds of new genes are generated from different origins of gene duplication. This study reveals that the second type organism is most suitable to generate many kinds of new genes and the third type organism is next in line. The cell differentiation is a representative character, which requires many kinds of genes for its expression, and the present result provides an explanation for the fact that the cell differentiation has started in the second type of organisms and then evolved to the higher hierarchy in the third type of organisms.

## 2. The concept of biological activity

Although the 'biological activity' is a macroscopic quantity generally characterizing various biological systems such as an ecological system, an organism, an individual cell of a multicellular organism etc. (Otsuka, 2004, 2005, 2008), it will be explained focusing on an organism for the present purpose of considering the large-scale evolution of organisms by gene duplication. In general, an organism may be characterized by a set of two macro-variables, the genome size  $N$  and its systematization -  $S_N$  of genes and their products. The systematization corresponds to the negentropy, which should be measured for the specific arrangement of nucleotides in individual genes, the degree of accuracy in transmitting the genetic information to the amino acid sequences of proteins, the formation of metabolic pathways by enzyme protein functions, the regulation and control at various levels of biological processes, the cell structure constructed by the interaction of metabolic products, and for furthering the communication between differentiated cells in the case of multicellular organisms. The energy acquired by an organism depends not only on the

genome size  $N$  and systematization -  $S_N$  but also on the material and energy source  $M$  available from the environment. Thus, the energy acquired by the organism during its lifetime is expressed as  $E_a(M; N, S_N)$ , which may be an increasing function of  $N$  and  $S_N$  as well as of  $M$ . On the other hand, the organism utilizes the acquired energy and materials to construct the biomolecules for its growth and self-reproduction. The energy  $E_s(N, S_N)$  stored in the form of biomolecules is also another increasing function of  $N$  and  $S_N$ . The difference between the acquired energy and the stored energy,  $E_a(M; N, S_N) - E_s(N, S_N)$ , is lost as heat. According to the second law of thermodynamics, the entropy production by the heat must compensate for the entropy reduction, i. e., -  $S_N$ , by the systematization. Thus, the following inequality must hold:

$$E_a(M; N, S_N) - E_s(N, S_N) - TS_N > 0 \quad (1)$$

where  $T$  is the temperature. In other words, this indicates the upper boundary of systematization (negentropy) by entropy production (Otsuka & Nozawa, 1998). However, organisms must have developed the systematization to increase the acquired energy through the evolutionary process of gene and genome duplication, nucleotide base substitutions and selection, and this is the main problem in the present study. The larger value of  $E_a(M; N, S_N) - E_s(N, S_N) - TS_N$  gives a measure for the biological processes to proceed more smoothly. In this sense, the quantity of  $E_a(M; N, S_N) - E_s(N, S_N) - TS_N$ , which an organism produces during one generation, will be called the 'biological activity' of the organism. The 'biological activity' has thermodynamic connotation as a departure from equilibrium, but this is in a reverse relation to the free energy in thermodynamics, which decreases upon any change in a given system by the decrease in internal energy and/or by the increase in entropy. In an organism, the acquired energy is stored in ATP and NADH molecules as chemical energy, and it is gradually consumed in the syntheses of biomolecules under the guidance of the enzymes, without drastically raising the temperature. In such moderate reactions, the temperature is almost constant, and the quantity obtained from the 'biological activity' divided by the product of the Boltzmann constant  $k$  and temperature  $T$  is considered to be approximately proportional to the self-reproducing rate of an organism, which will be denoted by  $R(M; N, S_N)$  hereafter.

This concept of 'biological activity' or self-reproducing rate is useful to formulate the large-scale of evolution arising from the gene duplication and succeeding generation of new genes. The essence of the present theory considers the following process of evolution in terms of 'biological activity'. First, the enlarged genome size  $N + \Delta N$  due to gene duplication makes the stored energy  $E_s(N + \Delta N, S_N)$  larger than  $E_s(N, S_N)$ , while the acquired energy  $E_a(M; N + \Delta N, S_N)$  remains almost equal to  $E_a(M; N, S_N)$ . Thus, the 'biological activity' of a variant bearing duplicated genes becomes lower than that of the original style organism. Moreover, the biological activity of the variant further decreases by the increase in systematization from  $S_N$  to  $S_{N+\Delta N}$ , as a new gene generated from the counterpart of duplicated genes is incorporated into an extended system of regulation and control. However, such a variant with the lower activity is not necessarily extinct but has a chance to recover as a new style of organisms, if the new gene begins expressing a new biological function to raise the acquired energy from  $E_a(M, N + \Delta N, S_N)$  to  $E_a(M', N + \Delta N, S_{N+\Delta N})$  by utilizing the new material and energy source  $M'$  other than  $M$ , or by moving to a new living area or by utilizing  $M$  more efficiently in the case of  $M' = M$ . This process of the large-scale evolution will be mathematically formulated to estimate the probabilities of generating new genes, for the

first type of organisms in section 3, for the second type in section 4 and for the third type in section 5.

### 3. Prokaryotes and lower eukaryotes in the monoploid state

For the mathematical description, the set of variables  $(N_i, S_{Ni})$  characterizing a variant  $i$  will be simply denoted as a single variable  $x_i$ , unless the description of changes in its content is necessary. In the population of monoploid organisms taking a common material and energy source  $M$ , the number  $n(x_i; t)$  of variants, each characterized by the monoploid genome  $x_i$ , obeys the following time-change equation.

$$\frac{d}{dt}n(x_i; t) = \{Q_{xi}(t)R(M; x_i) - D(x_i)\}n(x_i; t) + \sum_{j(\neq i)} q_{xi, xj}(t)R(M; x_j)n(x_j; t) \quad (2)$$

where the self-reproducing rate and death rate of the variant  $x_i$  are denoted by  $R(M; x_i)$  and  $D(x_i)$ , respectively. The apparent decrease factor  $Q_{xi}(t)$  in the self-reproducing rate of the variant  $x_i$  is related with the mutation term  $q_{xj, xi}(t)$  from the variant  $x_i$  to other kinds of variants  $x_j$ 's in the following way.

$$Q_{xi}(t) = 1 - \sum_{j(\neq i)} q_{xj, xi}(t) \quad (3)$$

If the quantity  $q_{xi, xi}(t)$  defined by  $Q_{xi}(t) - 1$  is introduced, the restriction  $j \neq i$  can be removed from the summation of the second term on the right side of Eq. (2). For investigating the population behavior, Eq. (2) is transformed into the following two types of equations; one concerning the total number of all kinds of variants defined by  $B(t) = \sum_i n(x_i; t)$  and another concerning the fraction  $f(x_i; t)$  of variants  $x_i$  defined by  $n(x_i; t)/B(t)$ .

$$\frac{d}{dt}B(t) = W_{av}(M; t)B(t) \quad (4)$$

$$\frac{d}{dt}f(x_i; t) = \{W(M; x_i) - W_{av}(M; t)\}f(x_i; t) + \sum_j q_{xi, xj}(t)R(M; x_j)f(x_j; t) \quad (5)$$

where the increase rate  $W(M; x_i)$  of variant  $x_i$  and the average increase rate  $W_{av}(M; t)$  of organisms in the population are defined by the following forms, respectively.

$$W(M; x_i) \equiv R(M; x_i) - D(x_i) \quad (6)$$

$$W_{av}(M; t) \equiv \sum_i W(M; x_i)f(x_i; t) \quad (7)$$

Strictly, the nucleotide base change occurs due to the miss in repairing damaged bases, while the gene duplication occurs by the illegitimate crossing over of DNA strands upon replication. Although they are simply represented by the mutation term  $q_{xi, xj}(t)$  in the above mathematical formulation, the point mutation due to nucleotide base change and the gene duplication are distinguished from each other in the following mathematical treatment.

Darwinian evolution corresponds to the evaluation of the time-change of variant fractions mainly by the first term on the right side of Eq. (5), as discussed by Eigen (1971). If the increase rate  $W(M;x_i)$  of an occasionally generated mutant  $x_i$  is greater than the average increase rate, that is,  $W(M;x_i) - W_{av}(M;t) > 0$ , the fraction  $f(x_i;t)$  increases with time according to the first term on the right side of Eq. (5). The increase in the fraction of such variants  $x_i$  gradually raises the average increase rate  $W_{av}(M;t)$ , resulting in the increase in the total number  $B(t)$  of organisms according to Eq. (4), although this increase is ultimately stopped by the decrease in available material  $M$ . On the other hand, the fraction  $f(x_i;t)$  decreases when  $W(M;x_i) - W_{av}(M;t) < 0$ . Thus, the organisms taking a common material and energy source  $M$  are elaborated by mutation and selection, and most of them finally reach the ones with the optimum increase rate, each characterized by  $x_{opt}$ . However, such Darwinian evolution may only hold for the point mutations in existing genes.

The large-scale evolutionary process of generating new gene(s) from gene duplication is obtained by evaluating the fraction of variants up to the first and higher orders of the mutation term. For this illustration, Eq. (5) will be formally integrated with respect to time  $t$ :

$$f(x_i;t) = \exp\left[\int_0^t \{W(M;x_i) - W_{av}(M;\tau)\} d\tau\right] \left[\int_0^t \sum_j q_{xi,xj}(\tau) R(M;x_j) f(x_j;\tau) \right. \\ \left. [-\int_0^\tau \{W(M;x_i) - W_{av}(M;\tau')\} d\tau'] d\tau + f(x_i;0)\right] \quad (8)$$

After the organisms  $x_{opt}$  have become dominant in the population,  $W_{av}(M;t)$  is approximately equal to  $W(M; x_{opt})$ , the fractions of variants except for  $x_{opt}$  are neglected on the right side of Eq. (8), and the mutation term  $q_{xi,xopt}(t)$  is replaced by the mutation rate  $q_{xi,xopt}$  defined as an average of mutation terms during a sufficiently long time  $t$ , i. e.,

$$q_{xi,xopt} \equiv \frac{1}{t} \int_0^t q_{xi,xopt}(\tau) d\tau \quad (9)$$

Then, the fraction  $f(x_i)$  of variants  $x_i$  is finally related with the fraction  $f(x_{opt})$  of dominant organisms  $x_{opt}$  in the following form.

$$f(x_i) = \frac{q_{xi,xopt} R(M;x_{opt})}{W(M;x_{opt}) - W(M;x_i)} f(x_{opt}) \quad (10)$$

Among such satellite variants, the variant arising from the gene duplication is especially notable in the sense that it has the potential to generate a new gene from the counterpart of duplicated genes. If the probability of generating a new gene  $I$  from the duplicated part in  $x_i$  is denoted by  $q_{xI,x_i}$ , a new style of the organism carrying the new gene  $I$  is generated from the original style of an organism with the following probability  $P_{m1}(x_I \leftarrow x_i \leftarrow x_o)$ .

$$P_{m1}(x_I \leftarrow x_i \leftarrow x_o) = \frac{q_{xI,x_i} q_{xi,xo} R(M;x_o)}{W(M;x_o) - W(M;x_i)} \quad (11)$$

where  $x_{opt}$  is rewritten into  $x_o$  with the meaning of the original style of an organism. Here,  $x_i$  and  $x_I$  correspond to  $(N+\Delta N, S_N)$  and  $(N+\Delta N, S_{N+\Delta N})$ , respectively, in terms of the set of variables characterizing an organism in section 2.

When a biologically meaningful character is newly exhibited by two new genes generated from different origins of gene duplication, the variant, which experienced gene duplication  $i$ , must successively experience further gene duplication  $j$  in the other part of the genome to exhibit such a new character. The fraction  $f(x_{ij}; t)$  of such variants  $x_{ij}$  obeys the following equation as a special case of Eq. (5).

$$\frac{d}{dt} f(x_{ij}; t) = \{W(M; x_{ij}) - W_{av}(M; t)\} f(x_{ij}; t) + q_{xij, xi}(t) R(M; x_i) f(x_i; t) \quad (12)$$

where  $q_{xij, xi}(t)$  represents the mutation term from the variant  $x_i$  to the variant  $x_{ij}$  and the smaller terms including the mutation from the variant  $x_{ij}$  to other variants are neglected. By formally integrating Eq. (12), the fraction  $f(x_{ij})$  of variants  $x_{ij}$  is finally expressed as

$$f(x_{ij}) = \frac{q_{xij, xi} R(M; x_i)}{W(M; x_{opt}) - W(M; x_{ij})} f(x_i) \quad (13)$$

where  $W_{av}(M; t)$  is approximated to be  $W(M; x_{opt})$  and the mutation term  $q_{xij, xi}(t)$  is replaced by the mutation rate  $q_{xij, xi}$ , i. e.,

$$q_{xij, xi} \equiv \frac{1}{t} \int_0^t q_{xij, xi}(\tau) d\tau \quad (14)$$

By inserting the expression (10) of fraction  $f(x_i)$  into the right side of Eq. (13), the fraction  $f(x_{ij})$  of variants  $x_{ij}$  is related with the fraction  $f(x_{opt})$  of dominant organisms  $x_{opt}$  by the second order of mutation rates in the following form.

$$f(x_{ij}) = \frac{q_{xij, xi} R(M; x_i)}{W(M; x_{opt}) - W(M; x_{ij})} \cdot \frac{q_{xi, xopt} R(M; x_{opt})}{W(M; x_{opt}) - W(M; x_i)} f(x_{opt}) \quad (15)$$

Thus, a new style of the organism  $x_{IJ}$  carrying new genes  $I$  and  $J$  is generated from the original style of an organism  $x_o$  with the following probability  $P_{m2}(x_{IJ} \leftarrow x_{ij} \leftarrow x_o)$ .

$$P_{m2}(x_{IJ} \leftarrow x_{ij} \leftarrow x_o) = \frac{q_{xIJ, xIJ} q_{xI, xi} q_{xij, xi} R(M; x_i)}{W(M; x_o) - W(M; x_{ij})} \cdot \frac{q_{xi, xo} R(M; x_o)}{W(M; x_o) - W(M; x_i)} \quad (16)$$

where  $q_{xIJ, xIJ}$  is the probability of generating the new gene  $J$  from the duplicated part in  $j$ . This procedure can be easily extended to the general case of successively generating three or more new genes.

Before describing the result of the general case, the expression of probabilities (11) and (16) will be simplified by assuming that the gene duplication only reduces the self-reproducing rate of the variant without any influence on the death rate. When the self-reproducing rate of the original style organism is simply denoted by  $R$  and that of the variant  $x_i$  is expressed as  $R(1-s_1)$  with the reduction factor satisfying  $0 < s_1 < 1$ , the probability (11) is simply expressed as

$$P_{m1}(x_I \leftarrow x_i \leftarrow x_o) = \frac{Q_1}{s_1} \quad (17)$$



where  $q_{x_l, x_i} q_{x_i, x_0}$  is denoted by  $Q_1$ . In the same way, the self-reproducing rate of the variant  $x_{ij}$  is denoted as  $R(1 - s_1 - s_2)$  with the additional reduction factor  $s_2$  under the condition of  $0 < s_1 + s_2 < 1$  and  $q_{x_{lj}, x_{lj}} q_{x_l, x_i} q_{x_{ij}, x_i} q_{x_i, x_0}$  is denoted by  $Q_2$ . The expression of the probability (16) then becomes

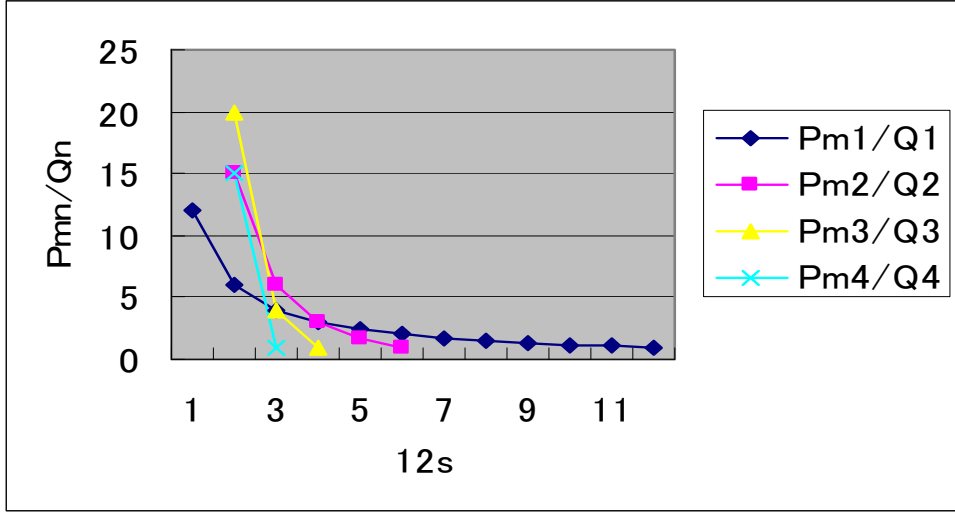


Fig. 1. The probabilities of generating new genes from gene duplication in the monoploid organism. On the basis of Eq. (20), the values of  $P_{mn}/Q_n$  are plotted against the twelve-fold reduction factor  $12s$  for  $n = 1, 2, 3$  and  $4$ . Although the value of  $Q_n$  becomes smaller for a larger value of  $n$ , the plotting of the probability  $P_{mn}$  in the unit of  $Q_n$  makes the figure compact. The probability  $P_{m1}$  is present in a whole range of reduction factor  $0 < s < 1$ . As the number of  $n$  increases, however, the range of reduction factor  $s$ , where the probability  $P_{mn}$  is present, is narrowed to  $0 < s < 1/n$ .

$$P_{m2}(x_{lj} \leftarrow x_{ij} \leftarrow x_o) = \frac{(1-s_1)}{s_1(s_1+s_2)} Q_2 \quad (18)$$

This expression of probabilities (17) and (18) is easily extended to express the probability of successively generating  $n$  kinds of new genes in the following way.

$$P_{mn} = \frac{(1-s_1)(1-s_1-s_2) \cdots (1-s_1-s_2-s_3-\cdots-s_{n-1})}{s_1(s_1+s_2) \cdots (s_1+s_2+s_3+\cdots+s_n)} Q_n \quad (19)$$

The reduction factors  $s'_i$ s in Eq. (19) are in the relations of  $0 < s_1 + s_2 + \cdots + s_n < 1$  and  $0 < s_1, s_2, \cdots, s_n < 1$ . Strictly, the values of  $s'_i$ s are different depending on the length of duplicated sequences and on the order of gene duplication events. For the simple investigation of the  $n$  dependence of  $P_{mn}$ , however, these reduction factors are assumed to be commonly equal to one variable  $s$ . Then, the first relation becomes  $0 < s < 1/n$ , and Eq. (19) is reduced to

$$P_{mn} = \frac{(1-s)(1-2s)(1-3s) \cdots \{1-(n-1)s\}}{n!s^n} Q_n \quad (20)$$

On the basis of this expression (20), the probabilities  $P_{mn}'$ s for several values of  $n$  are plotted against the reduction factor  $s$  in Fig. 1. In the case of  $n = 1$ , the reduction factor  $s$  is permitted in a whole range of  $0 < s < 1$  and the probability  $P_{m1}$  of generating a new gene is present in this range. This means that the monoploid organism is suitable to create a new gene step by step, testing the biological function of the new gene product, even if the gene size is large. As the value of  $n$  increases, however, the reduction factor  $s$  is restricted to the narrower range of  $0 < s < 1/n$ . When the monoploid organism creates simultaneously multiple kinds of new genes from different origins of gene duplication, therefore, these genes are obliged to be of a smaller size. Moreover, the probability  $P_{mn}$  is also decreased as the value of  $n$  increases. This is because  $Q_n$  becomes smaller for the larger value of  $n$ . Thus, it is difficult for the monoploid organism to evolve a new character which requires the expression of many kinds of new and large genes. This result is common to the prokaryote with a single DNA molecule and the lower eukaryote with the plural number of chromosomes, if the latter does not conjugate to exchange homologous chromosomes.

#### 4. The monoploid eukaryotes that exchange homologous chromosomes through conjugation

Some monoploid eukaryotes with the plural number of chromosomes conjugate to form a zygote during their life cycle, and the zygote produces monoploid descendants by exchanging homologous chromosomes upon the meiosis. Although the conjugation also occurs in prokaryotes, it only takes place to exchange plasmids and partial genes. Originally, the conjugation would have evolved to avoid the accumulation of disadvantageous mutations in a special lineage and to maintain the stability of a population by weakening the influence of such mutations. However, the conjugation in the eukaryote with the plural number of chromosomes makes it possible to produce the descendant receiving two or more new genes, even if these new genes are relatively large. Thus, the conjugation of such eukaryotes is considered to be the strategy to overcome the difficulty of generating many and large new genes from the successive gene duplication in a single lineage of monoploid organisms. For this illustration, several examples will be first listed in the following subsections 4.1 to 4.3, and they are used to estimate the probabilities of producing the descendant received more new genes by the conjugation of variants, each carrying a smaller number of new genes.

##### 4.1 The probability of producing the descendant received two new genes

Such a descendant is produced from the conjugation of two types of variants, one carrying a new gene  $I$  on a chromosome  $C_1$  and another carrying a new gene  $J$  on another kind of chromosome  $C_2$ . The genome of the variant carrying the new gene  $I$  is denoted by  $(C_{1l}, C_{20})$  and the genome of another variant carrying the new gene  $J$  is by  $(C_{10}, C_{2j})$ . The conjugation of these two types of variants yields the zygote, whose genome constitution is represented by  $(C_{1l}, C_{10}; C_{2j}, C_{20})$ . If the homologous chromosomes are randomly partitioned into two daughter cells, the probability  $P_{c2}$  of producing the new monoploid descendant received the genome  $(C_{1l}, C_{2j})$  is calculated to be  $P_{m1}^2/2$ .

##### 4.2 The probability of producing the descendant received three new genes

The descendant received three new genes  $I$ ,  $J$  and  $K$  can be produced from the conjugation of variants, one carrying one new gene  $I$  and another carrying two new genes  $J$  and  $K$ . Two cases are considerable for this production.

One is the case that the new gene  $I$  is encoded on the chromosome  $C_1$  and both new genes  $J$  and  $K$  are encoded on another kind of chromosome  $C_2$ . Then, the genome of the variant carrying the new gene  $I$  is denoted by  $(C_{1I}, C_{20})$  and the genome of another variant carrying the new genes  $J$  and  $K$  is denoted by  $(C_{10}, C_{2JK})$ . The conjugation of these two variants forms the zygote  $(C_{1I}, C_{10}; C_{2JK}, C_{20})$ , which can produce four types of monoploid descendants,  $(C_{1I}, C_{2JK})$ ,  $(C_{1I}, C_{20})$ ,  $(C_{10}, C_{2JK})$  and  $(C_{10}, C_{20})$ . If the homologous chromosomes are equivalently partitioned into two daughter cells, regardless of carrying new genes or not, the new monoploid descendant  $(C_{1I}, C_{2JK})$  is produced with the probability of  $P_{m1}P_{m2}/2$ .

In the second case, the new genes  $J$  and  $K$  are encoded on separate chromosomes. If the chromosome carrying the new gene  $K$  is denoted by  $C_3$ , the genome of the variant carrying new genes  $J$  and  $K$  is represented by  $(C_{10}, C_{2J}, C_{3K})$ . The conjugation of this variant and the variant  $(C_{1I}, C_{20}, C_{30})$  forms the zygote  $(C_{1I}, C_{10}; C_{2J}, C_{20}; C_{3K}, C_{30})$ . Under the random partition of homologous chromosomes, this zygote yields a new monoploid descendant  $(C_{1I}, C_{2J}, C_{3K})$  with the probability of  $P_{m1}P_{m2}/4$ .

As a whole,  $3P_{m1}P_{m2}/4$  is obtained for the probability  $P_{c3}$  of producing a new monoploid organism received three new genes by conjugation.

### 4.3 The probability of producing the descendant received four new genes

The highest probability of producing the descendant received four new genes is obtained by the conjugation of two variants, one carrying two new genes  $I$  and  $J$ , and another carrying other two new genes  $K$  and  $L$ . The following three cases (i) ~ (iii) are considerable. (i) The new genes  $I$  and  $J$  are encoded on the chromosome  $C_1$  in one variant, while the new genes  $K$  and  $L$  are encoded on the chromosome  $C_2$  in another variant. The conjugation of these two variants forms the zygote  $(C_{1IJ}, C_{10}; C_{2KL}, C_{20})$ , which yields four types of monoploid descendants,  $(C_{1IJ}, C_{2KL})$ ,  $(C_{1IJ}, C_{20})$ ,  $(C_{10}, C_{2KL})$  and  $(C_{10}, C_{20})$ . If the homologous chromosomes are randomly partitioned into two descendants, the probability of producing the monoploid descendant  $(C_{1IJ}, C_{2KL})$  is calculated to be  $P_{m2}^2/2$ . (ii) The new genes  $I$  and  $J$  are encoded on the chromosome  $C_1$  in one variant but the new genes  $K$  and  $L$  are encoded on the chromosomes  $C_2$  and  $C_3$ , respectively, in another variant. The conjugation of these two variants forms the zygote  $(C_{1IJ}, C_{10}; C_{2K}, C_{20}; C_{3L}, C_{30})$ . If the homologous chromosomes in each kind of 1, 2 and 3 are randomly partitioned into two daughter cells, the probability of producing the monoploid descendant  $(C_{1IJ}, C_{2K}, C_{3L})$  is calculated to be  $P_{m2}^2/4$ . (iii) The new genes  $I$  and  $J$  are encoded on the chromosomes  $C_1$  and  $C_2$ , respectively, in one variant, while the new genes  $K$  and  $L$  are encoded on the chromosomes  $C_3$  and  $C_4$ , respectively, in another variant. The conjugation of these two variants forms the zygote  $(C_{1I}, C_{10}; C_{2J}, C_{20}; C_{3K}, C_{30}; C_{4L}, C_{40})$ , and yields the monoploid descendant  $(C_{1I}, C_{2J}, C_{3K}, C_{4L})$  with the probability  $P_{m2}^2/8$ .

The monoploid organism receiving four new genes can be also produced by the conjugation of a variant with one new gene  $I$  on the chromosome  $C_1$  and another variant with three new genes  $J$ ,  $K$  and  $L$ . The following three cases (iv) ~ (vi) are considerable for the location of the three new genes  $J$ ,  $K$  and  $L$ . (iv) The three new genes are encoded on the same chromosome  $C_2$ . In this case, the conjugation of the two variants forms the zygote  $(C_{1I}, C_{10}; C_{2JKL}, C_{20})$  and yields the monoploid descendant  $(C_{1I}, C_{2JKL})$  with the probability of  $P_{m1}P_{m3}/2$ . (v) The new gene  $J$  is encoded on the chromosome  $C_2$  and the other two new genes  $K$  and  $L$  are encoded on the chromosome  $C_3$ . The conjugation of these variants forms the zygote  $(C_{1I}, C_{10}; C_{2J}, C_{20}; C_{3KL}, C_{30})$  and yields the descendant monoploid  $(C_{1I}, C_{2J}, C_{3KL})$  with the probability of

$P_{m1}P_{m3}/4$ . (vi) The new genes  $J$ ,  $K$  and  $L$  are encoded on the chromosomes  $C_2$ ,  $C_3$  and  $C_4$ , respectively. In this case, the probability of producing the monoploid descendant ( $C_{1l}$ ,  $C_{2l}$ ,  $C_{3k}$ ,  $C_{4l}$ ) is further decreased to be  $P_{m1}P_{m3}/8$ .

As illustrated in the above examples in subsections 4.1 to 4.3, the probability  $P_{c2n}$  of producing the monoploid descendant received the even number  $2n$  of new genes through one time of conjugation is generally expressed as

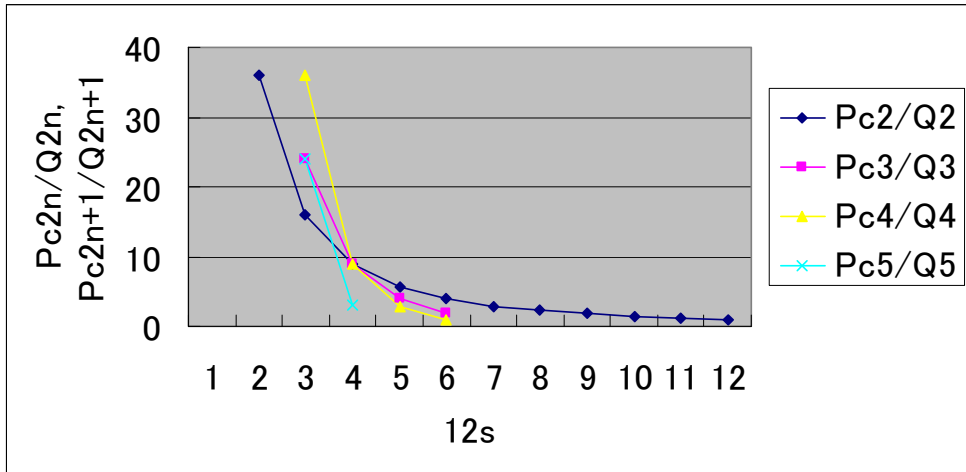


Fig. 2. The probabilities of producing the descendants received multiple kinds of new genes by the conjugation of monoploid organisms. The probability  $P_{c2n}$  of producing the descendant received  $2n$  kinds of new genes is simply expressed as the square of the probability  $P_{mn}$ , i. e.,  $P_{c2n} = P_{mn}^2$ . In the same way, the probability  $P_{c2n+1}$  of producing the descendant received  $(2n+1)$  kinds of new genes is expressed as the product of the probabilities  $P_{mn+1}$  and  $P_{mn}$ , i. e.,  $P_{c2n+1} = P_{mn+1}P_{mn}$ . Using the relations of  $Q_{2n} = Q_n^2$  and  $Q_{2n+1} = Q_nQ_{n+1}$ ,  $P_{c2n}/Q_{2n}$  and  $P_{c2n+1}/Q_{2n+1}$  are plotted against the twelve-fold reduction factor  $12s$  for  $n = 1$  and 2. It should be noted that the probabilities  $P_{c2n}$  and  $P_{c2n+1}$  are present in the wider range of reduction factor than the probabilities  $P_{m2n}$  and  $P_{m2n+1}$  shown in Fig. 1, respectively.

$$P_{c2n} = a_{n,n}P_{mn}^2 + b_{n+1,n-1}P_{mn+1}P_{mn-1} + \dots \quad (21)$$

and the probability  $P_{c2n+1}$  of producing the monoploid descendant received the odd number  $(2n+1)$  of new genes is expressed as

$$P_{c2n+1} = a_{n,n+1}P_{mn}P_{mn+1} + b_{n+2,n-1}P_{mn+2}P_{mn-1} + \dots \quad (22)$$

Although the coefficients  $a_{n,n}$ ,  $a_{n,n+1}$ ,  $b_{n+1,n-1}$ ,  $b_{n+2,n-1}$  etc. depend not only on the number of new genes but also on the distribution of new genes over chromosomes in a complex way, the first terms are most important on the right sides of Eqs. (21) and (22), respectively. This is because the probabilities  $P_{mn}$  and  $P_{mn+1}$  in these terms are present in the wider range of reduction factor than those in other terms, as indicated in the preceding section. Thus,  $P_{c2} \sim P_{m1}^2$  and  $P_{c2n+1} \sim P_{mn}P_{mn+1}$ , without the coefficients  $a_{n,n}$  and  $a_{n,n+1}$ , are plotted against the

reduction factor  $s$  for some values of  $n$  in Fig. 2. The probability  $P_{c2n}$  is present in the same range of reduction factor as the probability  $P_{mn}$  is present and the probability  $P_{c2n+1}$  is present in the same range of reduction factor as the probability  $P_{mn+1}$  is. This indicates that the larger size of new genes not generated from the successive gene duplication in a single lineage of monoploid organisms can be assembled into an organism through conjugation. Although the values of  $P_{c2n}$  and  $P_{c2n+1}$  are smaller than those of  $P_{mn}$  and  $P_{mn+1}$ , respectively, due to the relations of  $Q_{2n} < Q_n$  and  $Q_{2n+1} < Q_{n+1}$ , the smaller value of the probability only means the longer time for the monoploid organism to receive  $2n$  or  $(2n+1)$  new genes through the conjugation of variants than the time for a single lineage of monoploid organisms to generate  $n$  or  $(n+1)$  new genes from gene duplication. If these larger new genes assembled by conjugation endow the descendant with a superior new character, such descendants increase their fraction as a new style of organisms. In this sense, it should be also noted that the descendant receiving  $n$  ( $= 3, 4, 5, \dots$ ) kinds of new genes can be produced with the lower probability of  $(1/2)^{n(n-1)/2} P_{m1}^n$  by the successive conjugation of variants having experienced gene duplication on different kinds of chromosomes. Such successive hybridization of different variants, each of them carrying one new gene, may become the main course to yield a new style of organisms carrying three or more new genes, if the homologous chromosomes different in carrying two or more new genes, such as those appeared in the first case of subsection 4.2 and in (i), (ii), (iv) and (v) of subsection 4.3, are severely incompatible upon the meiosis in the zygote.

At any rate, the eukaryote with the plural number of chromosomes is suitable to create new characters each expressed by many kinds of new genes, through the conjugation exchanging homologous chromosomes. This explains the diversity of various living styles of eukaryotes, ranging from the unicellular organisms called the *Protoctista* evolving various intracellular organs to the multicellular organisms evolving cell differentiation. As will be discussed in the last section, it is evident from the phylogeny of eukaryotes that the multicellularity and cell differentiation have also started in the monoploid eukaryotes, although the higher hierarchy of cell differentiation has developed in the diploid eukaryotes.

## 5. Higher eukaryotes in the diploid state

The higher eukaryote in the diploid state is characterized by the pairs of homologous chromosomes, and its large-scale evolution contains the process to establish the homozygote of new genes as well as their generation from gene duplication. Although the number of homologous chromosome pairs is different depending on the species of diploid organisms, a specific pair of homologous chromosomes  $(x_i, x_k)$  will be first focused for simplicity, where the suffixes  $i$  and  $k$  denote different mutations on the respective chromosomes. The number  $n(x_i, x_k; t)$  of variants carrying such a pair  $(x_i, x_k)$  obeys the following time-change equation in the population of organisms exchanging the homologous chromosomes upon reproduction.

$$\begin{aligned} \frac{d}{dt} n(x_i, x_k; t) = & \sum_{j,l} Q(x_i, x_k; t)_{ijkl} R(M; x_i, x_k)_{ijkl} n(x_i, x_j; t) n(x_k, x_l; t) - D(x_i, x_k) n(x_i, x_k; t) \\ & + \sum_{i',k'(\neq i,k)} \sum_{j,l} q(x_i, x_k \leftarrow x_{i'}, x_{k'}; t)_{i'j \times k'l} R(M; x_{i'}, x_{k'})_{i'j \times k'l} n(x_{i'}, x_j; t) n(x_{k'}, x_l; t) \end{aligned} \quad (23)$$

where  $R(M; x_i, x_k)_{ijxkl}$  is the rate of producing the children  $(x_i, x_k)$  from the mating of a variant  $(x_i, x_j)$  with another variant  $(x_k, x_l)$  under a common material and energy source  $M$ , and  $D(x_i, x_k)$  is the death rate of the organism  $(x_i, x_k)$ . The apparent decrease factor  $Q(x_i, x_k; t)_{ijxkl}$  is related with the mutation term  $q(x_i, x_k' \leftarrow x_i, x_k; t)_{ijxkl}$  in the following way.

$$Q(x_i, x_k; t)_{ijxkl} = 1 - \sum_{i', k' (\neq i, k)} q(x_i, x_k' \leftarrow x_i, x_k; t)_{ijxkl} \quad (24)$$

Although Eq. (23) makes no distinction between the male and the female for simplicity, this distinction does not essentially alter the following process of evolution.

In the same way as for monoploid organisms, the population behavior of diploid organisms becomes transparent by transforming Eq. (23) into the equation concerning the total number of organisms given by  $B(t) = \sum_i \sum_k n(x_i, x_k; t)$  and that concerning the fraction of variants  $(x_i, x_k)$  defined by  $f(x_i, x_k; t) = n(x_i, x_k; t)/B(t)$ . These equations are expressed in the following forms, respectively.

$$\frac{d}{dt} B(t) = \overline{W}(t) B(t) \quad (25)$$

$$\begin{aligned} \frac{d}{dt} f(x_i, x_k; t) &= \{W(x_i, x_k; t) - \overline{W}(t)\} f(x_i, x_k; t) \\ &+ \sum_{i', k'} \sum_{j, l} q(x_i, x_k \leftarrow x_{i'}, x_{k'}; t)_{i' j x k' l} R(M; x_{i'}, x_{k'})_{i' j x k' l} f(x_{i'}, x_j; t) f(x_{k'}, x_l; t) B(t) \end{aligned} \quad (26)$$

where the increase rate  $W(x_i, x_k; t)$  of the variant  $(x_i, x_k)$  is defined by

$$W(x_i, x_k; t) \equiv \sum_j \sum_l R(M; x_i, x_k)_{ijxkl} f(x_i, x_j; t) f(x_k, x_l; t) B(t) / f(x_i, x_k; t) - D(x_i, x_k) \quad (27)$$

the average increase rate  $\overline{W}(t)$  is by

$$\overline{W}(t) \equiv \sum_i \sum_k W(x_i, x_k; t) f(x_i, x_k; t) \quad (28)$$

and  $q(x_i, x_k \leftarrow x_i, x_k'; t)_{ijxkl}$  is defined by  $Q(x_i, x_k'; t)_{ijxkl} - 1$ . If the suffixes  $i, j, k$  and  $l$  denote the point mutations in existing genes, Eq. (26) represents Darwinian evolution gradually leading to the organisms with an optimal increase rate, each characterized by  $(x_{opt}, x_{opt})$ .

Because the gene duplication occurs only rarely, it is natural to consider that the large-scale evolution due to gene duplication starts after the organisms  $(x_{opt}, x_{opt})$  have been dominant in the population. If the chromosome having experienced gene duplication is newly denoted by  $x_i$  and the point mutation is neglected, the fraction  $f(x_i, x_{opt}; t)$  of variants  $(x_i, x_{opt})$  obeys the following equation as a special case of Eq. (26).

$$\begin{aligned} \frac{d}{dt} f(x_i, x_{opt}; t) &= \{W(x_i, x_{opt}; t) - \overline{W}(t)\} f(x_i, x_{opt}; t) \\ &+ q(x_i, x_{opt} \leftarrow x_{opt}, x_{opt}; t)_{optoptoptoptopt} R(M; x_{opt}, x_{opt})_{optoptoptoptopt} f^2(x_{opt}, x_{opt}; t) B(t) \end{aligned} \quad (29)$$

where the increase rate  $W(x_i, x_{opt}; t)$  of the variant  $(x_i, x_{opt})$  is given by

$$W(x_i, x_{opt}; t) = R(M; x_i, x_{opt})_{ioptxoptopt} f(x_{opt}, x_{opt}; t) B(t) - D(x_i, x_{opt}) \quad (30)$$

and the average increase rate  $\overline{W}(t)$  is by

$$\overline{W}(t) = W(x_i, x_{opt}; t) f(x_i, x_{opt}; t) + W(x_{opt}, x_{opt}; t) f(x_{opt}, x_{opt}; t) \quad (31)$$

The probability of generating a new style of organisms carrying a new gene is derived from Eq. (29). In the population where the organisms  $(x_{opt}, x_{opt})$  are dominant,  $\overline{W}(t)$  is approximately equal to  $W(x_{opt}, x_{opt})$ , and both  $f(x_{opt}, x_{opt}; t)$  and  $B(t)$  are hardly dependent on time. Eq. (29) is then integrated to give the following relation between the fraction of variants  $f(x_i, x_{opt})$  and that of dominant organisms  $f(x_{opt}, x_{opt})$ .

$$f(x_i, x_{opt}) = \frac{q(x_i, x_{opt} \leftarrow x_{opt}, x_{opt})_{optoptxoptopt} R(M; x_{opt}, x_{opt})_{optoptxoptopt} f^2(x_{opt}, x_{opt}) B}{W(x_{opt}, x_{opt}) - W(x_i, x_{opt})} \quad (32)$$

where the rate of generating the gene duplication  $i$  is defined for a sufficiently long time  $t$  by

$$q(x_i, x_{opt} \leftarrow x_{opt}, x_{opt})_{optoptxoptopt} \equiv \frac{1}{t} \int_0^t q(x_i, x_{opt} \leftarrow x_{opt}, x_{opt}; \tau)_{optoptxoptopt} d\tau \quad (33)$$

Although this relation (32) seems to be different in including the population size  $B$  from Eq. (10) of monoploid organisms at first glance, the denominator on the right side of Eq. (32) also contains the population size  $B$  as seen in Eqs. (27) and (30). If the population size is large enough to neglect the difference in death rate between the variant  $(x_i, x_{opt})$  and the dominant organism  $(x_{opt}, x_{opt})$ , therefore, the difference in the increase rate  $W(x_{opt}, x_{opt}) - W(x_i, x_{opt})$  is approximately equal to  $\{R(M; x_{opt}, x_{opt})_{optoptxoptopt} - R(M; x_i, x_{opt})_{ioptxoptopt}\} B f(x_{opt}, x_{opt})$ , and Eq. (32) is reduced to be

$$f(x_i, x_{opt}) = \frac{q(x_i, x_{opt} \leftarrow x_{opt}, x_{opt})_{optoptxoptopt} R(M; x_{opt}, x_{opt})_{optoptxoptopt} f(x_{opt}, x_{opt})}{R(M; x_{opt}, x_{opt})_{optoptxoptopt} - R(M; x_i, x_{opt})_{ioptxoptopt}} \quad (34)$$

This is essentially the same form as Eq. (10) of the monoploid organism in the case when the gene duplication hardly changes the death rate, i. e.,  $D(x_i) \approx D(x_{opt})$ . Denoting the probability of generating a new gene  $I$  from the gene duplicated part  $i$  by  $q(x_i \leftarrow x_i)$ , the probability  $P_{d1}(x_i, x_o \leftarrow x_o, x_o)$  that a new style of the organism  $(x_i, x_o)$  carrying the new gene  $I$  heterogeneously is generated from the original style of an organism  $(x_o, x_o)$  is expressed as

$$P_{d1}(x_i, x_o \leftarrow x_i, x_o) = \frac{q(x_i \leftarrow x_i) q(x_i, x_o \leftarrow x_o, x_o)_{ooxxoo} R(M; x_o, x_o)_{ooxxoo}}{R(M; x_o, x_o)_{ooxxoo} - R(M; x_i, x_o)_{ioxxoo}} \quad (35)$$

where  $x_{opt}$  in Eq. (34) is rewritten into  $x_o$  with the meaning of the original type chromosome. Thus, a new style diploid organism also arises from the minor members in the population just like the case of monoploid organisms.

However, the content of the above probability in diploid organisms is different from the case of monoploid organisms in the following points. First of all, the reproducing rate  $R(M; x_i, x_o)_{ioxxoo}$  is only the half of  $R(M; x_o, x_o)_{ooxxoo}$  even in the random partition of homologous chromosomes, and the former may be further decreased by the lowering of the biological activity of the variant  $(x_i, x_o)$ . Second, the further gene duplication to produce two or more new genes is hardly expected in the homologous chromosomes  $(x_i, x_o)$ , because the fraction

of such variants experienced successive gene duplication becomes much lower, not only due to the severer lowering of biological activity but also by the severer incompatibility of homologous chromosomes or by the separation of the chromosomes carrying different origins of duplicated genes in the descendants. That is, if the further gene duplication  $j$  occurs on the chromosome  $x_i$  to yield  $x_{ij}$ , for example, the incompatibility of chromosomes  $x_{ij}$  and  $x_o$  becomes severer upon the mitosis and/or the meiosis. If the gene duplication  $j$  occurs on the chromosome  $x_o$  to yield  $x_j$ , on the contrary, the chromosome  $x_j$  is separated from the chromosome  $x_i$  in the descendants.

In spite of such conservative property, the diploid organism with the plural number of homologous chromosome pairs can give rise to a new style of an organism getting together two or more new genes, through the successive hybridization among the satellite variants having experienced gene duplication on different kinds of chromosomes. As the first example, the appearance of a new style organism received two kinds of new genes  $I$  and  $J$  will be considered by this mechanism of hybridization. In this case, two pairs of homologous chromosomes  $(x_o, x_o; y_o, y_o)$  are focused, and the probability of generating the heterozygote  $(x_i, x_o; y_j, y_o)$  from the original style of organisms  $(x_o, x_o; y_o, y_o)$  is considered through the hybridization of two types of variants  $(x_i, x_o; y_o, y_o)$  and  $(x_o, x_o; y_j, y_o)$ . According to Eq. (35), this probability  $P_{d2}(x_i, x_o; y_j, y_o \leftarrow x_o, x_o; y_o, y_o)$  is given by

$$\begin{aligned} & P_{d2}(x_i, x_o; y_j, y_o \leftarrow x_o, x_o; y_o, y_o) \\ &= \frac{q(x_i \leftarrow x_o)q(x_j, y_o \leftarrow x_o, x_o)_{0000 \times 0000} R(M; x_o, x_o; y_o, y_o)_{0000 \times 0000}}{R(M; x_o, x_o; y_o, y_o)_{0000 \times 0000} - R(M; x_i, x_o; y_o, y_o)_{i000 \times 0000}} \\ & \quad \cdot \frac{q(y_j \leftarrow y_o)q(y_j, y_o \leftarrow y_o, y_o)_{0000 \times 0000} R(M; x_o, x_o; y_o, y_o)_{0000 \times 0000}}{R(M; x_o, x_o; y_o, y_o)_{0000 \times 0000} - R(M; x_o, x_o; y_j, y_o)_{00j0 \times 0000}} r_2 \end{aligned} \quad (36)$$

where  $r_2$  is the ratio of the children received two kinds of new genes  $I$  and  $J$ , taking the value of  $(1/2)^2$  in the case of random partition of homologous chromosomes. In order to show the result of further hybridization process, Eqs. (35) and (36) will be simplified in their expression at this stage. The probabilities  $q(x_i \leftarrow x_o)$  and  $q(x_j \leftarrow x_o)$  of generating new genes  $I$  and  $J$  from duplicated parts  $i$  and  $j$  in Eqs. (35) and (36) may be equal to the corresponding probabilities  $q_{xI,xi}$  and  $q_{xJ,xIj}$  in Eqs. (11) and (16), respectively, because the nucleotide base substitution rate is almost common to both eukaryotes and prokaryotes (Kimura, 1980; Otsuka et al., 1997). Although it is still difficult to estimate the occurrence frequency of gene duplication, this frequency is also assumed to be common to both monoploid and diploid organisms, i. e.,  $q_{xi,xo} \sim q(x_i, x_o \leftarrow x_o, x_o)_{0000 \times 0000}$  and  $q_{xij,xi} \sim q(y_j, y_o \leftarrow y_o, y_o)_{0000 \times 0000}$ , for simplicity. The reproducing rates  $R(M; x_o, x_o; y_o, y_o)_{0000 \times 0000}$ ,  $R(M; x_i, x_o; y_o, y_o)_{i000 \times 0000}$  and  $R(M; x_o, x_o; y_j, y_o)_{00j0 \times 0000}$  are simply denoted by  $R$ ,  $R(1 - S_1)$  and  $R(1 - S_2)$ , respectively, with the reduction factors  $S_1$  and  $S_2$ , where both  $S_1$  and  $S_2$  satisfy the relation  $1/2 < S_1, S_2 < 1$  as noted already. Eqs. (35) and (36) are then rewritten into

$$P_{d1}(x_i, x_o \leftarrow x_o, x_o) = \frac{Q_1}{S_1} \quad (37)$$

and

$$P_{d2}(x_i, x_o; y_j, y_o \leftarrow x_o, x_o; y_o, y_o) = \frac{Q_2}{S_1 S_2} r_2 \quad (38)$$



respectively. Here,  $Q_1$  and  $Q_2$  represent the terms  $q(x_1 \leftarrow x_i)q(x_i, x_o \leftarrow x_o, x_o)_{00x00}$  and  $q(x_1 \leftarrow x_i)q(x_i, x_o \leftarrow x_o, x_o)_{0000x0000}q(y_1 \leftarrow y_j)q(y_j, y_o \leftarrow y_o, y_o)_{0000x0000}$ , respectively. As the extension, the probability  $P_{dn}$ , with which a new style diploid organism carrying  $n$  kinds of new genes heterogeneously is generated from the successive hybridization of variants, is expressed in the following form.

$$P_{dn} = \frac{Q_n}{S_1 S_2 \dots S_n} r_n \quad (39)$$

where  $S_i$  ( $i=1, 2, \dots, n$ ) is the reduction factor in the producing rate of the variant carrying duplicated genes on the chromosome  $i$ ,  $Q_n$  is the product of the probabilities of generating  $n$  kinds of new genes from gene duplication on the respective chromosomes and  $r_n$  is the ratio of the children received these new genes. Although reduction factors  $S_i$ 's in Eq. (39) independently take values in the range of  $1/2 < S_i < 1$ , they are tentatively represented by a common variable  $S$  for a simple illustration of  $n$  dependence of  $P_{dn}$  in a figure. Then, the probability  $P_{dn}$  is simply expressed by

$$P_{dn} = \frac{Q_n}{S^n} r_n \quad (40)$$

These probabilities  $P_{dn}$ 's in Eq. (40) are plotted against the reduction factor  $S$  in Fig. 3 for several values of  $n$ . As noted already, the reduction factor  $S$  is restricted to the

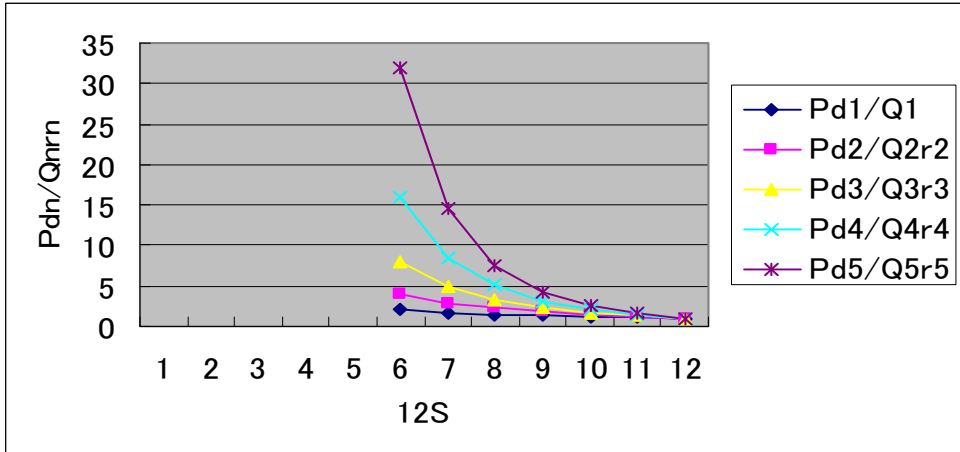


Fig. 3. The probabilities of generating new genes from gene duplication and successive hybridization in diploid organisms. On the basis of Eq. (40), the values of  $P_{dn}/Q_n r_n$  are plotted against the twelve-fold reduction factor  $12S$  for  $n = 1, 2, 3, 4$  and  $5$ . The value of  $r_1$  is equal to one, and the curve of  $P_{d1}$  vs  $S$  is consistent with the curve of  $P_{m1}$  vs  $s$  in Fig. 1, but the range of reduction factor  $S$  is restricted to the range of  $1/2 < S < 1$ . For a larger value of  $n$ , however, the probability  $P_{dn}$  is still present in the range of  $1/2 < S < 1$ . Although  $P_{dn+1}/Q_{n+1} r_{n+1}$  is larger than  $P_{dn}/Q_n r_n$  in the figure,  $P_{dn+1}$  is smaller than  $P_{dn}$ . This is because  $Q_{n+1} r_{n+1}$  is smaller than  $Q_n r_n$ , as discussed in the text.

range of  $1/2 < S < 1$ , and the probability  $P_{dn}$  is within the range of  $2^n Q_n r_n > P_{dn} > Q_n r_n$ . If the homologous chromosomes are randomly partitioned into the children regardless of carrying a new gene or not,  $r_n$  takes the value of  $(1/2)^n$  and the above relation of  $P_{dn}$  becomes  $Q_n > P_{dn} > Q_n/2^n$ . Moreover, the value of  $Q_n$  becomes smaller for the larger value of  $n$ , and the probability  $P_{dn}$  becomes lower as the number of new genes assembled by hybridization is increased. The lower probability means the longer time or more generations for a new style organism carrying more kinds of new genes to appear. Thus, the diploid organism has a chance to acquire many kinds of new genes by hybridization, but it takes a longer time to realize this chance.

Moreover, the process to establish the homozygote is further continued after the new style organism carrying  $n$  kinds of new genes heterogeneously is generated with the probability  $P_{dn}$ . Although it is laborious to follow this process completely, the essence of this process can be elucidated by investigating the ratio of children that receive these new genes homogeneously and heterogeneously from the mating between the organisms each carrying  $n$  kinds of new genes heterogeneously. If the chromosomes in each homologous pair are randomly partitioned into the children regardless of carrying a new gene or not, the ratio of children receiving  $(n-k)$  kinds of new genes is calculated to be  ${}_nC_k 3^{n-k}/4^n$  with the normalization factor  $4^n$ , where  $k$  takes a value ranging from zero to  $n$ . This indicates that more than half of the children receive all new genes ( $k = 0$ ) for  $n = 1, 2$ . If the one or two new genes exhibit an excellent character, therefore, the descendants increase their fraction monotonously as a new style of organisms. However, the ratio of children receiving a full set of new genes becomes smaller for a larger value of  $n$ . In the case of  $n = 5$ , for example, the ratio of the children that receive five kinds of new genes ( $k = 0$ ) decreases to  $(3/4)^5$ , while other five types of children each appear with the ratio of  $(3/4)^4/4$  by receiving four kinds of new genes ( $k = 1$ ) in different ways. When a biologically meaningful character is expressed by five kinds of new genes, therefore, only  $(3/4)^5$  of the children succeed in expressing this character but other five types of children are reserved as those carrying 'hidden genes' for producing other characters by further hybridization with other types of variants. Such divergence of characters becomes more outstanding when a larger number of new genes are required for the expression of a character. This divergent property in the process to establish many kinds of new genes as the homozygote explains the explosive divergence of body plans that has occasionally occurred in diploid organisms, because the cell differentiation is a representative character expressed by many kinds of genes and its hierarchical evolution constructs body plans, as will be discussed in the next section. Until the new style organisms are established as the homozygote, the mating between the variants of heterozygote also regenerates the original style of organisms. The phenomenon called the "reversion" or "atavism" in classical biology may be the vestige of this evolutionary process to establish the homozygote.

If the influence of transposons is explicitly considered, it makes the above process more complicated in such a way that duplicated genes are separately transferred to different kinds of chromosomes. When various origins of duplicated genes or new genes are concentrated on one chromosome, however, the descendants received such a chromosome may be extinct due to the incompatibility of this chromosome with its partner chromosome not carrying any new gene. Thus, many kinds of new genes for expressing a new character may be scattered over different kinds of chromosomes in survivors just like the result of the present model scheme.

## 6. Conclusions and discussion

The variants, which experienced gene duplication, first decline to be minor members in a population by the load of carrying extra gene(s), but some of them revives as a new style of organisms by the generation of new gene(s) from the counterpart of duplicated genes. After the new gene(s) appear, the new style organisms increase their fraction being further elaborated by Darwinian evolution. This course of the large-scale evolution is essentially the same in any type of organisms, and this is a necessary condition for the new style of organisms and the original style of organisms to be able to coexist utilizing different material and energy sources or to live in separate areas, showing a striking contrast to the survival of the fittest in Darwinian evolution. This evolutionary pattern also gives an explanation to the punctuated mode of evolution, which has been proposed from paleontology against the gradual accumulation of variants in Darwinian evolution (Eldredge & Gould, 1972).

However, the detailed processes of this large-scale evolution are different depending on the types of genome constitution and transmission. The monoploid organism is suitable to generate one new gene step by step testing its biological function, but hardly generates many kinds of new genes simultaneously. The lower eukaryote, whose genome consists of the plural number of chromosomes, resolves this difficulty to produce a new style of organisms receiving many kinds of new genes by the conjugation of variants carrying different origins of new genes. The diploid organism can also produce a new character responsible for multiple kinds of new genes by the successive hybridization of different variants but its conservative property requires the succeeding process to establish the homozygote of these genes. This process becomes longer for a larger number of new genes to be established. During this long process, the further hybridization with other variants also occurs, occasionally yielding the explosive divergence of new characters depending on the combinatorial sets of new genes. This conclusion of the present study explains the recently revealed evolutionary patterns of prokaryotes and eukaryotes to a great extent, getting an insight into the problems how and why the monoploid eukaryotes have evolved to the diploid eukaryotes.

According to the analyses of base-pair changes in ribosomal RNAs, the main lineages of present-day prokaryotes diverged  $3.0 \times 10^9$  years ago, developing various chemical syntheses,  $O_2$ -releasing photosynthesis and  $O_2$  respiration, respectively (Otsuka et al., 1999), after the earlier divergence of archaeobacteria, eubacteria and eukaryotes (Sugaya & Otsuka, 2002). Several stages from simple electron transport pathways to  $O_2$  respiration and  $O_2$ -releasing photosynthesis are still observed in the present-day eubacteria and the elongation of the pathways has taken place stepwise by gene duplication, as can be traced from the amino acid sequence similarities between their component proteins and the ubiquitous permeases (Otsuka, 2002; Otsuka & Kawai, 2006), although such similarity search of amino acid sequences is not systematically carried out yet for chemical syntheses. However, the excellent abilities of  $O_2$  respiration and  $O_2$ -releasing photosynthesis cannot be fully exhibited in the simple cell structure of prokaryotes (Otsuka, 2005), and the genome size of the eubacteria having these abilities is also limited to the order of  $10^6$  bp compactly encoding 3,000 ~ 4,000 genes like the other prokaryotes (Wheeler et al., 2004).

On the other hand, the eukaryotes have experienced much more evolutionary events until some of them establish the diploid state. The ancestral eukaryote probably became the predator of eubacteria by developing the intracellular structure, endocytosis and exocytosis

as well as the signal transduction network. Such cell structure would have been suitable to acquire the mitochondria as the endosymbionts of  $O_2$ -respiratory eubacteria, which is estimated to have occurred  $2.0 \times 10^9$  years ago (Margulis, 1981; Yang et al., 1985; Otsuka et al., 1999). Under the supply of abundant ATP molecules efficiently synthesized by the mitochondria, the yeast *Sacchromyces cerevisiae*, which appeared  $1.8 \times 10^9$  years ago (Otsuka et al., 1997), has expanded its genome to  $1.2 \times 10^7$  bp encoding 6,300 genes (Wheeler et al., 2004) and can take the diploid state under nutrient conditions, although it usually takes the monoploid state. The enlarged genome consisting of the plural number of chromosomes also requires the special apparatus for the faithful segregation of sister DNAs upon cell division, in contrast to the prokaryotes where the membrane attachment mechanism of DNA only operates (Jacob et al., 1963; Ogden et al., 1988; de Boer, 1993). Multiple kinds of gene products such as the primitive spindle pole and kinetochore and polar microtubules are already present in the yeast (Alberts et al., 1994) while several bundles of microtubules only pass through tunnels in the typical *Dinoflagellates* (Kubai, 1975; Hearth, 1980; Wise, 1988). Thus, the components of this auxiliary apparatus for cell division may have evolved step by step at the stage of unicellular eukaryotes. Although the molecular mechanism underlying the switching from the monoploid to diploid states and vice versa is not fully clarified yet, the example of yeast indicates that this mechanism itself has also evolved at the stage of unicellular eukaryotes.

However, the evolution from the monoploid eukaryote to the diploid eukaryote has taken place considerably gradually via several stages. This is reasonable because the diploid state is an extreme case of gene duplication. If the genome size jumps from  $N$  to  $2N$ , this means the increase in the stored energy and systematization from  $E_s(N, S_N)$  and  $S_N$  to  $E_s(2N, S_{2N})$  and  $S_{2N}$ . Thus, the acquired energy must be also increased to maintain the biological activity. As indicated already (Otsuka, 2008), this increase in acquired energy is possibly attained by the cooperative action of differentiated cells. However, the evolution of cell differentiation cannot occur suddenly. On this problem, the present result throws light, in the point that the conjugation of lower eukaryotes with the plural number of chromosomes is suitable to assemble many kinds of new genes necessary for cell differentiation. In fact, the recently revealed phylogeny of eukaryotes strongly suggests at least the following five stages in the evolution from the monoploid to diploid eukaryotes. (a) First, the monoploid eukaryote evolves the conjugation to exchange the homologous chromosomes. (b) Second, this eukaryote then develops multicellularity and cell differentiation in the monoploid state by assembling many kinds of new genes. (c) Third, the cell differentiation also advances to the cells in the diploid state. (d) Fourth, the eukaryote evolves to alternate the monoploid generation and the diploid generation. (e) Finally, the eukaryote evolves to the diploid organism with the higher hierarchy of cell differentiation.

As far as the present knowledge of the phylogeny of eukaryotes and their genome constitution (Otsuka et al., 1997) is concerned, the first lineages having evolved multicellularity and cell differentiation are some of the fungi that appear after yeast and the sea algae, which have further acquired photosynthetic eubacteria as the endosymbionts in the lineage of fungi (Van den Eynde et al., 1988). However, the most advanced one of them still remains at the stage (d), alternating the monoploid generation and the diploid generation. Apart from the lineages of fungi and algae, the evolution of advancing the cell differentiation to the diploid state has taken place in the animals and the green plants, whose divergence is estimated to have occurred  $1.2 \times 10^9$  years ago (Dickerson, 1971). Among

them, the green plants, which have also acquired the chloroplasts as the endosymbionts of photosynthetic eubacteria independently of sea algae, provide a representative example of the above five stages of evolution from the monoploid organisms to the diploid organisms. The *Cojugatae* such as *Roya* and *Spirogyra* are at the stage (a), the *Chara* of *Charophyta* is at the stage (b), the *Bryophyta* is at the stage (c), where the fertilized egg on female gametophyte grows into sporangium, the *Pterophyta* is at the stage (d), and the seed plants are at the stage (e). According to the recent analysis of neutral nucleotide base substitutions in *rbcL* genes on the chloroplast genomes (Kawai & Otsuka, 2004), the divergence of *Charophyta* and *Bryophyta* occurred more than  $10^9$  years ago, the divergence of *Bryophyta* and *Pterophyta* occurred around  $4.7 \times 10^8$  years ago, and the divergence of *Pterophyta* and seed plants occurred about  $3.8 \times 10^8$  years ago.

The molecular mechanism underlying the cell differentiation is not fully clarified yet, but it is probably based on a set of receptors, the corresponding ligands, signal transduction proteins, transcriptional regulators as well as the proteins exhibiting the respective cell-type specific functions. Moreover, the amino acid sequences of these proteins under the control of signal transduction network become longer by the attachment of special amino acid residue repeats such as serines and threonines. Thus, the assembly of so many kinds of large genes into a genome must have first progressed under the scheme of the conjugation of monoploid eukaryotes with the plural number of chromosomes. After a set of genes responsible for cell differentiation are established in the monoploid state, the increase in the repertoire of the respective members would have occurred relatively easily. In particular, a small number of nucleotide base substitutions could bring about the expansion of such protein families as transcriptional regulators, receptors and kinases associated with the signal transduction network, although these kinds of proteins have their origins at the stage of unicellular eukaryotes. The increase in acquired energy by the cell differentiation in the monoploid state makes it possible to realize the cell differentiation in the diploid state. The example of green plants suggests that the cell differentiation in the diploid state has started from the zygote and gradually spread to form other organs of diploid cells, resulting in the alternation of the monoploid generation and the diploid generation. The diploid state is suitable to protect the differentiated cells from the point mutations, as will be discussed in the last part of this section, but it takes a longer time or many generations to establish a set of many genes for advancing the further cell differentiation in the diploid state as the homozygote. Although this is the barrier lying between the stage (d) and the stage (e), the diploid organisms having gone over this barrier receive a good chance to produce various combinatorial sets of new genes leading to the explosive divergence of morphological characters. Such explosive divergence has the merit of testing simultaneously various characters for survival.

Although any example of animals at the stages (b) and (c) is hardly found at the present time, the *Cnidaria* still alternates the monoploid generation and the diploid one. The divergence of *Cnidaria* and the common ancestor of other animals occurred immediately after the animal-plant divergence (Otsuka & Sugaya, 2003). The famous explosion of body plans giving rise to *Annelida*, *Mollusca*, *Arthropoda*, *Echinodermata* and *Chordata*, which is first found by the fossil record of Ediacara and Avalon faunas (Mathews & Missarzshersky, 1975; Rozanov & Zhuravlev, 1992) and of Cambrian Burgess Shale (Gould, 1989) and then estimated to have occurred successively during the period of  $9 \sim 6 \times 10^8$  years ago by the analysis of neutral nucleotide base substitutions (Otsuka & Sugaya, 2003), is probably based on the evolutionary scheme of diploid organisms described in sections 5, because these

animals show the living style defined as the diploid organism in the present chapter. Such divergence of body plans occasionally occurred afterwards in each of the above phyla. The examples well investigated in paleontology are the divergence of *Placodermi*, cartilaginous fish and bony fish, the divergence of amphibians, reptiles and mammals, and the divergence of dinosaurs and birds, which occurred in the *Chordata* within the recent  $4 \times 10^8$  years (Carroll, 1988). The seed plants also show the similar tendency in the successive divergence of *Coniferophyta*, *Anthophyta* and their relatives (Fairon-Demaret & Scheckler, 1987; Rothwell et al., 1989; Rowe, 1992; Stewart & Rothwell, 1993; Kawai & Otsuka, 2004), although many of these seed plants can also self-reproduce by the parthenogenesis and their explosive feature seems mild. Although the explosive divergence of body plans can be also explained by the biological activity expressed in terms of the interaction between differentiated cells (Otsuka, 2008), the present study derives this divergence from the aspect of the generation of new genes from gene duplication in diploid organisms.

The fossil record of these examples indicates that the original style of organisms prospered over a wide region when new styles of organisms diverged, being consistent with the present theory. The prosperity of the original style of organisms means that their biological activity is high, and this is necessary to permit the existence of variants carrying duplicated genes in the population and further to enhance the chance of assembling many kinds of new genes into a genome by hybridization. This is in contrast with Darwinian evolution generating new species adapted to the special environment of a geographically isolated district by accumulating point mutations.

Finally, some discussions will be given to the problem why the cell differentiation has been shifted from the monoploid state to the diploid state. This problem arises from the present result that the diploid organism is not necessarily superior to the monoploid organism with the ability of exchanging homologous chromosomes in assembling many kinds of new genes for cell differentiation. The main reason of this shifting may be the protection of differentiated cells from the point mutations due to the miss in repairing damaged nucleotide bases. First of all, many more genes are needed to develop the higher hierarchy of cell differentiation. In fact, the genome size of higher eukaryotes is expanded to the order of  $10^8 \sim 10^9$  bp, e. g.,  $1.2 \times 10^8$  bp encoding 24,000 genes in *Arabidopsis thaliana*,  $1.4 \times 10^8$  bp encoding 13,000 genes in *Drosophila melanogaster* and  $3.1 \times 10^9$  bp encoding 30,000 genes in *Homo sapiens* (Wheeler et al., 2004). Second, it takes a longer time, one or more years, to develop the higher hierarchy of cell differentiation to form an adult form in the higher eukaryotes, although the growth rate and the lifetime seem to be further regulated differently depending on species. On the other hand, the mutation rate due to the miss in repair is  $10^{-9}$  per site per year in eukaryotes as well as in prokaryotes (Kimura, 1980; Otsuka et al., 1997). As the evidence for the above discussion, the males of some species of ants and bees are born by the haploid parthenogenesis, showing that the monoploid state is sufficient for the high hierarchy of cell differentiation during their short lifetime. Although the accuracy in repairing damaged DNAs can be raised by the additional energy for proofreading (Hopfield, 1974), the evolution of organisms has not been directed to use such additional energy. On the contrary, the nucleotide base substitution rate becomes about tenfold faster in animal mitochondrial genome than in the host cell genome, as is used to resolve the phylogeny of recently diverged animals (Hasegawa et al., 1985; Pesole et al., 1999; Otsuka et al., 2001). This faster mutation rate strongly suggests that the energy to proofread the small genome of mitochondria is diminished and instead the saved energy is used to raise the biological activity of the host cell. For the

same sequence length of gene duplication, therefore, the reduction factor may take a smaller value in animals than in lower eukaryotes and prokaryotes. Thus, the fraction of variants carrying the 'hidden genes' generated from gene duplication may be high enough to hybridize between them in higher eukaryotes, especially in animals. Such 'hidden genes' belong to the category of 'genetic polymorphism', which has been first proposed by Ford (1965) and is subsequently disclosed by electrophoretic studies, although the 'genetic polymorphism' was only regarded as the result of random fixation of selectively neutral or nearly neutral mutations by the neutralist (Kimura, 1977).

It is still somewhat mysterious that the introns and spacers are more expanded in animal genomes than in the genomes of other eukaryotes. Such expansion can be seen from the ratio of the genome size to the number of encoded genes described above. It is conceivable that the introns are necessary for messenger RNAs to pass through the nuclear membrane and the spacers enhance the crossing over of homologous chromosomes without injuring established genes, but the expansion of introns and spacers in the higher eukaryotes might imply any other biological role of their nucleotide sequences.

## 7. References

- Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K. & Watson, J. D. (1994). *Molecular Biology of the Cell*. 3<sup>rd</sup> Ed. Garland Publishing Inc., New York & London. pp. 941-943
- Birney, E. & Other fifty persons (Ensembl 2006). Database Issue. *Nucleic Acids Res.* Vol. 1, D556-D561
- Carroll, R. L. (1988). *Vertebrate Paleontology and Evolution*. W. H. Freeman, New York
- Darwin, C. (1859). *The Origin of Species*. John Murry, London
- de Boer, R. A. J. (1993). Chromosome Segregation and Cytokinesis in Bacteria. *Curr. Opin. Cell Biol.*, Vol 5, pp. 232-237
- Dickerson, R. E. (1971). The Structure of Cytochrome c and the Rate of Evolution. *J. Mol. Evol.*, Vol. 1, pp. 26-45
- Dobzhansky, T. (1941). *Genetics and the Origin of Species*. 2<sup>nd</sup> Ed. Columbia University Press, New York
- Eigen, E. (1971). Selforganization of Matter and Evolution of Biological Macromolecules. *Die Naturwissenschaften*, Vol. 58, pp. 465-523
- Eldredge, N. & Gould, S. J. (1972). Punctuated Equilibria: An Alternative to Phyletic Gradualism. In: *Models in Paleobiology*, T. J. M. Schopf, (Ed.). p. 82, Freeman and Cooper, San Francisco
- Fairon-Demaret, M. & Scheckler, E. S. (1987). Typification and Redescription of *Moresnetia zalesskyi* Stockmans, 1948, an Early Seed Plant from the Upper Famennian of Belgium. *Bull Inst Roy Sci Nat Belg Sci Terre*, Vol. 57, pp. 193-199
- Ferris, S. D. & Whitt, G. S. (1979). Evolution of the Differential Regulation of Duplicated Gene after Polyploidization. *J. Mol. Evol.*, Vol. 12, pp. 267-317
- Fisher, R. A. (1930). *The General Theory of Natural Selection*. Oxford Univ. Press, London and New York
- Ford, E. B. (1965). *Genetic Polymorphism*. Faber & Faber, London
- Gilbert, W. (1978). Why Genes in Pieces ? *Nature*, Vol. 271, p. 501
- Gould, S. J. (1989). *Wonderful Life. The Burgess Shale and the Nature of History*. W. W. Norton & Company Inc., New York

- Hasegawa, M.; Kishino, H. & Yano, T. (1985). Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA. *J. Mol. Evol.*, Vol. 22, pp. 160-174
- Hearth, I. B. (1980). Variant Mitosis in Lower Eukaryotes: Indicators of the Evolution of Mitosis? *Int. Rev. Cytol.*, Vol. 64, pp. 1-80
- Hopfield, J. J. (1974). Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity. *Proc. Nat. Acad. Sci. USA*, Vol. 71, pp. 4135-4139
- Huxley, J. (1943). *Evolution: The Modern Synthesis. The Grandson of T. H. Huxley Explores Mendelism, Evolutionary Trends, and Genetic Systems*. Harper & Row, New York
- Ingram, V. M. (1963). *The Hemoglobin in Genetics and Evolution*. Columbia Press, New York
- Jacob, F.; Brenner, S. & Cuzin, F. (1963). On the Regulation of DNA Replication in Bacteria. *Cold Spr. Harb. Symp. Quant. Biol.*, Vol. 28, pp. 329-348
- Kawai, Y. & Otsuka, J. (2004). The Deep Phylogeny of Land Plants Inferred from a Full Analysis of Nucleotide Base Changes in Terms of Mutation and Selection. *J. Mol. Evol.*, Vol. 58, pp. 479-489
- Kimura, M. (1977). Causes of Evolution and Polymorphism at the Molecular Level. *Proceedings of the Second Taniguchi International Symposium on Biophysics*, pp. 1-28, Mishima, Japan
- Kimura, M. (1980). A Simple Method for Estimating Evolutionary Rate of Base Substitutions through Comparative Studies of Nucleotide Sequences. *J. Mol. Evol.*, Vol. 16, pp. 111-120
- Kojima, S. & Otsuka, J. (2000a). Characterization of Organisms by the Paralogous Relationships of Proteins. Part I. *Escherichia coli*. *Res. Commun. in Biochemi Cell & Molec. Biology*, Vol. 4, pp. 59-82
- Kojima, S. & Otsuka, J. (2000b). Characterization of Organisms by the Paralogous Relationships of Proteins. Part II. *Methanococcus jannaschii*. *Res. Commun. in Biochemi. Cell & Molec. Biology*, Vol. 4, pp. 83-100
- Kojima, S. & Otsuka, J. (2000c). Characterization of Organisms by the Paralogous Relationships of Proteins. Part III. *Saccharomyces cerevisiae*. *Res. Commun. in Biochemi. Cell & Molec. Biology*, Vol. 4, pp. 101-138
- Kojima, S. & Otsuka, J. (2002). Characterization of Proteome by Similarity Linkages of Paralogous Functional Domains and Special Amino Acid-Rich Regions. Part IV. *Drosophila melanogaster*. *Res. Commun. in Biochemi. Cell & Molec. Biology*, Vol. 6, pp. 72-102
- Kubai, D. F. (1975). The Evolution of the Mitotic Spindle. *Int. Rev. Cytol.*, Vol. 43, pp. 167-227
- Margulis, L. (1981). *Symbiosis in Cell Evolution: Life and its Environment on the Early Earth*. W. H. Freeman, San Francisco
- Mathews, S. C. & Missarzhevsky, V. (1975). Small Shelly Fossils of Late Precambrian and Early Cambrian Age: A Review of Recent Work. *Q. J. Geol. Soc. London*, Vol. 131, pp. 289-304
- Mayer, E. (1942). *Systematics and the Origin of Species. A Correlation of the Evidence and Points of View of Systematics With Those of Other Biological Disciplines, Particularly Genetics and Ecology*. Columbia University Press, New York
- Nowak, M. A.; Bonhoeffer, S. & May, R. M. (1994). Spatial Games and the Maintenance of Cooperation. *Proc. Natl. Acad. Sci. USA*, Vol. 91, pp. 4877-4881



- Ogden, G. B.; Pratt, M. J. & Schaechter, M. (1988). The Replication Origin of the *Escherichia coli* Chromosome Binds to Cell Membrane only When Hemimethylated. *Cell*, Vol. 54, pp. 127-135
- Ohno, S. (1970). *Evolution by Gene Duplication*. Spring-Verlag, Berlin.
- Otsuka, J.; Nakano, T. & Terai, G. (1997). A Theoretical Study on the Nucleotide Changes under a Definite Functional Constraint of Forming Stable Base-Pairs in the Stem Regions of Ribosomal RNAs; Its Application to the Phylogeny of Eukaryotes. *J. Theor. Biol.*, Vol. 184, pp. 171-186
- Otsuka, J. & Nozawa, Y. (1998). Self-Reproducing System can Behave as Maxwell's Demon: Theoretical Illustration under Prebiotic Conditions. *J. Theor. Biol.*, Vol. 194, pp. 205-221
- Otsuka, J.; Terai, G. & Nakano, T. (1999). Phylogeny of Organisms Investigated by the Base-Pair Changes in the Stem Regions of Small and Large Ribosomal Subunit RNAs. *J. Mol. Evol.*, Vol. 48, pp. 218-235
- Otsuka, J.; Kawai, Y. & Sugaya, N. (2001). The Influence of Selection on the Evolutionary Distance Estimated from the Base Changes between Homologous Nucleotide Sequences. *J. Theor. Biol.*, Vol. 213, pp. 129-144
- Otsuka, J. (2002). An Inquiry into the Evolutionary History of Photosynthetic and Respiratory Systems from the Similarity Relationships of Member Proteins. In: *Recent Research Developments in Proteins*, Transworld Research Network, (Ed.), Vol. 1, pp. 229-256, Kerala, India
- Otsuka, J. & Sugaya, N. (2003). Advanced Formulation of Base Pair Changes in the Stem Regions of Ribosomal RNAs; Its Application to Mitochondrial rRNAs for Resolving the Phylogeny of Animals. *J. Theor. Biol.*, Vol. 222, pp. 447-460
- Otsuka, J. (2004). A Theoretical Characterization of Ecological System by Circular Flow of Materials. *Ecological Complexity*, Vol. 1, pp. 237-252
- Otsuka, J. (2005). A Theoretical Scheme for the Large-Scale Evolution of Organisms towards a Higher Order of Organization and Diversity. In: *Recent Research Developments in Experimental & Theoretical Biology*, Transworld Research Network, (Ed.), Vol. 1, pp. 93-122, Kerala, India
- Otsuka, J. & Kawai, Y. (2006). Phylogenetical Relationships among Permeases and the Membrane Proteins in Photosynthetic and Respiratory Systems. *Trends in Photochemistry and Photobiology*, Vol. 11, pp. 1-22
- Otsuka, J. (2008). A Theoretical Approach to the Large-Scale Evolution of Multicellularity and Cell Differentiation. *J. Theor. Biol.*, Vol. 255, pp. 129-136
- Pesole, G.; Gissi, C.; Chirico, A. D. & Saccone, C. (1999). Nucleotide Substitution Rate of Mammalian Mitochondrial Genomes. *J. Mol. Evol.*, Vol. 44, pp. 427-434
- Rothwell, G. W.; Scheckler, S. E. & Gillespie, W. H. (1989). *Elkinsia* gen. nov., a Late Devonian Gymnosperm with Cupulate Ovules. *Bot Gazette*, Vol. 150, pp. 170-189
- Rowe, N. P. (1992). Winged Late Devonian Seeds. *Nature*, Vol. 359, p. 682
- Rozanov, A. Y. & Zhuravlev, A. Y. (1992). The Lower Cambrian Fossil Record of the Soviet Union, In: *Origin and Early Evolution of the Metazoa*. J. H. Lipps & P. W. Signor, (Eds.), pp. 205-282, Plenum Press, New York and London
- Simpson, G. G. (1944). *Tempo and Mode in Evolution: A Synthesis of Paleontology and Genetics*, Columbia University Press, New York

- Stewart, N. S. & Rothwell, G. W. (1993). *Paleobotany and the Evolution of Plants*. pp. 438-467, Cambridge University Press, Cambridge
- Sugaya, N. & Otsuka, J. (2002). The Lineage-Specific Base-Pair Contents in the Stem Regions of Ribosomal RNAs and Their Influence on the Estimation of Evolutionary Distances. *J. Mol. Evol.*, Vol. 55, pp. 584-594
- Van den Eynde, H.; De Baere, R.; De Roeck, E.; Van de Peer, Y.; Vandenberghe, A.; Willekens, P. & De Wachter, R. (1988). The 5S Ribosomal RNA Sequences of a Red Algal Rhodoplast and Gymnosperm Chloroplast: Implication for the Evolution of Plastids and Cyanobacteria. *J. Mol. Evol.*, Vol. 27, pp. 126-132
- Wheeler, D. L.; Church, D. M.; Edgar, R.; Federhen, S.; Helmberg, W.; Madden, T. L.; Pontius, J. U.; Schuler, G. D.; Schriml, L. M.; Sequeira, E.; Suzek, T. O.; Tatusova, T. A. & Wagner, L. (2004). National Center of Biotechnology Information. *Nucl. Acid Res.*, Vol. 32, Database Issue D35
- Wise, D. M. (1988). The Diversity of Mitosis: the Value of Evolutionary Experiments. *Biochem. Cell Biol.*, Vol. 66, pp. 515-529.
- Wright, S. (1949). Adaptation and Selection. In: *Genetics, Paleontology and Evolution*. G. L. Jepson; G. G. Simpson & E. Mayer, (Eds.), pp. 365-389, Princeton Univ. Press, Princeton, New Jersey
- Yang, D.; Oyaizu, Y.; Oyaizu, H.; Olsen, G. J. & Woese, C. R. (1985). Mitochondrial Origin. *Proc. Natl. Acad. Sci. USA*, Vol. 82, pp. 4443-4447

# Duplicated Gene Evolution Following Whole-Genome Duplication in Teleost Fish

Baocheng Guo<sup>1,2,3</sup>, Andreas Wagner<sup>1,2</sup> and Shunping He<sup>3\*</sup>

<sup>1</sup> *Institute of Evolutionary Biology and Environmental Studies,  
University of Zurich, Zurich*

<sup>2</sup> *The Swiss Institute of Bioinformatics, Quartier Sorge-Batiment Genopode, Lausanne*

<sup>3</sup> *Fish Phylogenetics and Biogeography Group, Key Laboratory of Aquatic Biodiversity and Conservation, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan*

<sup>1,2</sup> *Switzerland*

<sup>3</sup> *PR China*

## 1. Introduction

Gene and genome duplication have been thought to play an import part during evolution since the 1930s (Bridges 1936; Stephens 1951; Ohno 1970) . Ohno (1970) proposed that the increased complexity and genome size of vertebrates has resulted from two rounds (2R) of whole genome duplication (WGD) in early vertebrate evolution, which provided raw materials for the evolutionary diversification of vertebrates. Recent genomic sequence data provide substantial evidence for the abundance of duplicated genes in many organisms. Extensive comparative genomics studies have demonstrated that teleost fish experienced another round of genome duplication, the so-called fish-specific genome duplication (FSGD) (Amores et al. 1998; Taylor et al. 2003; Meyer and Van de Peer 2005). Because the timing of this WGD and the radiation of teleost species approximately coincided, it has been suggested that the large number (about 27,000 species—more than half of all vertebrate species (Nelson, 2006)) of teleosts and their tremendous morphological diversity might be causally related to the FSGD event (Amores et al. 1998; Taylor et al. 2001; Taylor et al. 2003; Christoffels et al. 2004; Hoegg et al. 2004; Vandepoele et al. 2004). Semon and Wolfe (2007) showed thousands of genes that remained duplicated When Tetraodon and zebrafish diverged underwent reciprocal loss subsequently in these two species may have been associated with reproductive isolation between teleosts and eventually contributed to teleost diversification. A study in yeast demonstrated that speciation of polyploid yeasts may be associated with reciprocal gene loss at duplicated loci (Scannell et al. 2006). Thus, speciation accompanied by differential retention and loss of duplicated genes after genome duplication may be a powerful lineage-splitting force (Lynch and Conery 2000).

For two reasons, teleost fish represent an excellent model system to study the retention and loss of duplicated genes as well as their evolutionary trajectory following whole-genome

---

\* Corresponding author

duplication. First, many duplicated genes that resulted from the FSGD event were preserved in teleost genomes. Second, five teleost genomes have been sequenced and more teleost genomes are being sequenced. Here, we investigate retention, loss, and molecular evolution of duplicate genes after the FSGD in five available teleost genomes that include the genomes of zebrafish *Danio rerio*, stickleback *Gasterosteus aculeatus*, medaka *Oryzias latipes*, Takifugu *Takifugu rubripes*, and Tetraodon *Tetraodon nigroviridis*.

## 2. Identifying duplicated genes that resulted from the FSGD event throughout the teleost genomes

We obtained 23,155 gene families from the database HOMOLENS version 4 (<ftp://pbil.univ-lyon1.fr/databases/homolens4.php>) (Penel et al. 2009), which is based on the Ensembl release 49. We chose HOMOLENS, because it allowed us to reliably retrieve sets of orthologous genes for our evolutionary analysis. HOMOLENS is devoted to metazoan genomes from Ensembl and contains gene families from complete animal genomes found in Ensembl. HOMOLENS has the same architecture as HOVERGEN (Duret et al. 1994), in which genes are organized in families and include precalculated alignments and phylogenies. In HOMOLENS 4, alignments are computed using MUSCLE (Edgar 2004) with default parameters; phylogenetic trees are computed with PHYLML, using the JTT amino acid substitution model (Jones et al. 1992). Phylogenies are computed based on conserved blocks of the alignments selected with Gblocks (Castresana 2000). Each phylogenetic tree is reconciled with a species tree using the program RAP (Dufayard et al. 2005), which, combined with the tree pattern search functionality, allows detection of ancient gene duplications or selection of orthologous genes (Penel et al. 2009). Several studies on duplicated gene evolution have been performed with data retrieved from HOMOLENS (Brunet et al. 2006; Studer et al. 2008).

We employed a topology-based method to identify duplicated genes that resulted from the FSGD event in the five teleost genomes we study. Briefly, if two teleosts have been subject to the same whole genome duplication event, a gene X that has been duplicated in this event and retained in both genomes, should form two gene lineages “Xa” and “Xb” (Figure 1A). We identified gene trees with the topology shown in Figure 1A using the TreePattern functionality (Dufayard et al. 2005) of the FamFetch client for HOMOLENS. We required duplicated genes to exist in at least two species to increase the likelihood that they result from the FSGD event (Figure 1B). In total, we identified 1,500 gene families with duplicated genes in this way.

## 3. Differential retention and loss of duplicated genes during teleost diversification

The most common fate of a duplicated gene is nonfunctionalization (pseudogenization). After a whole genome duplication event, many genes share this fate, so that a genome's gene content may only appear be slightly increased long after the duplication (Wolfe and Shields 1997; Jaillon et al. 2004). Our data suggest that only 3.3 percent (zebrafish) to 7.2 percent (Takifugu) of genes in current teleost genomes result from the FSGD event (Table 1). These percentages are lower than the 13 percent of retained duplicates in yeast (Wolfe and Shields 1997). One possible reason for this difference might lie in our topology-based method to identify likely FSGD duplicates (Figure 1), which enforces duplicated genes to exist in at least

(A)

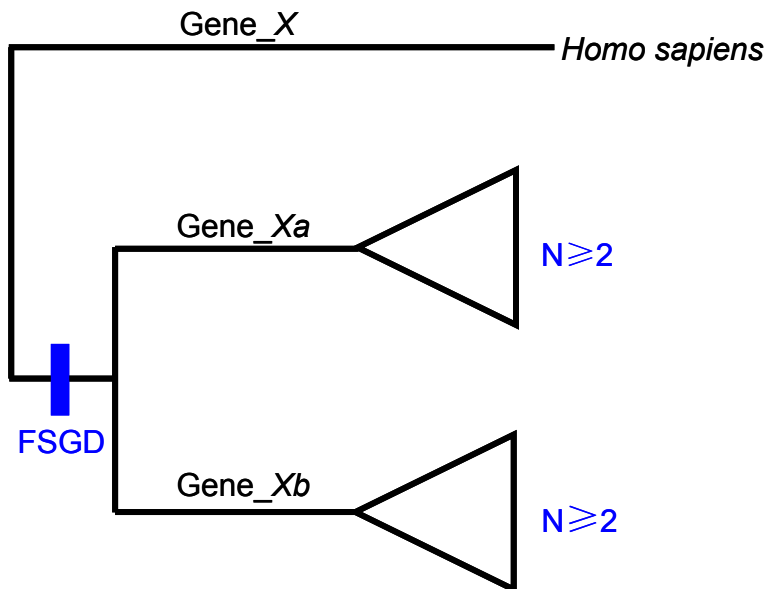


Fig. 1. (A) Expected phylogenetic relationship of duplicated gene *Xa* and *Xb* in two related species A and B when speciation occurred after the duplication event; (B) Tree topology we used for duplicated gene identification in the database HOMLENS 4. 'N ≥ 2' means that duplicated gene pairs must exist at least in two species to increase the likelihood that the duplicated genes actually resulted from the FSGD event.

	Number of genes *	Gene families with likely FSGD duplicates		
		FSGD Duplicates	Singleton	Double loss
<i>D. rerio</i>	21,420	731	541	228
<i>G. aculeatus</i>	20,839	681	669	150
<i>O. latipes</i>	19,687	1,162	311	27
<i>T. rubripes</i>	18,709	1,340	148	12
<i>Te. nigroviridis</i>	27,991	1,047	397	56

\* Total gene number in each genome, data based on the Ensembl release 49.

Table 1. Summary of different gene retention and loss in the 1,500 duplicated gene families we identified.

fish genomes. The FSGD occurred between 253 and 404 Million years ago (MYA) (Hoegg et al. 2004; Vandepoele et al. 2004), whereas the yeast whole genome duplication may have occurred more recently, between 100 and 150 MYA (Sugino and Innan 2005). More time has elapsed since the FSGD, allowing more duplicate genes to be lost.

Differential retention and loss of duplicated genes is a common phenomenon during speciation after genome duplication. It has been observed in yeast (Scannell et al. 2006) as well as in teleosts (Semon and Wolfe 2007), and is believed to lead to speciation. We thus expected that our dataset would contain many gene families with differential gene retention and loss, as well as fewer families where both copies are retained in all five teleost genomes. Indeed, when we consider all five species together, we observed that 90.4 percent of the 1,500 gene families we identified show differential retention and loss of duplicated genes, and in only 9.6 percent (144 gene families) are both copies retained in all five teleost genomes. Figure 2 and Table 1 show relevant data, broken down by study species. In 45.4 percent to 89.3 percent (depending on the species) of the 1,500 gene families we identified, both duplicates were retained. In 9.9 percent to 44.6 percent of the duplicates (depending on the species), one copy was lost. Our data also indicate that differences in differential gene retention are associated with the phylogenetic position and the relatedness between two teleost species (Figure 2). Taken together, these observations indicate that differential duplicated gene retention and loss are pervasive in teleosts, that the loss of duplicated genes is an ongoing process that has continued for hundreds of million years after the FSGD event, and that this process may be associated with teleost diversification.

We next discuss an illustrative example of differential duplicate gene retention and loss. It involves *Hox* genes, which encode a subclass of homeodomain transcription factors that help determine the anterior-posterior axis of bilaterian animals (McGinnis and Krumlauf 1992). In vertebrates, *Hox* genes have evolved a highly compact organization, where genes are arranged in clusters on chromosomes. *Hox* gene clusters are one of the best-studied systems for assessing gene retention and loss after the FSGD event (Amores et al. 1998; Prohaska and Stadler 2004; Hoegg et al. 2007; Guo et al. 2010), due to their genomic architecture and gene complement variation in teleosts. Seven or eight *Hox* clusters with different complements of

*Hox* genes exist in extant diploid teleosts. They are a result of the FSGD event, which was followed by loss of some *Hox* gene duplicates. The putative *Hox* cluster complement of the teleost ancestor and the *Hox* clusters of several model teleost species are shown in Figure 3. *Hox* clusters exhibit remarkably different gene complements in different teleost lineages after the FSGD event. Theoretically, 8 *Hox* clusters containing at least 80 *Hox* genes may have existed in the ancestor of teleosts after the FSGD event. Up to now, 66 of these *Hox* genes have been found in different teleost species and extant evolutionary diploid teleost usually have 45 to 49 *Hox* genes in their genome (Figure 3). According to the summary of Hoegg et al (2007) (Figure 3), the Ostariophysii have lost seven *Hox* genes since their hypothetical common ancestor with the Neoteleosts; during the evolution of the Neoteleosts eight *Hox* genes were lost; and the pufferfish lineage lost three genes in the common lineage leading to Takifugu and Tetraodon. Some *Hox* genes are specifically preserved in different teleosts, for example, *HoxA1b* has been identified thus far only in the Japanese eel (Guo et al. 2010). At the cluster level, eight *Hox* clusters were retained in basal species such as the Japanese eel (Guo et al. 2010) and the goldeye (Chambers et al. 2009), whereas one *Hox* cluster (C or D) was lost respectively in the Otocephala (Amores et al. 1998) and Euteleostei (Kurosawa et al. 2006). Based on the phylogeny of teleosts, Guo et al. (2010) proposed that the *HoxDb* cluster

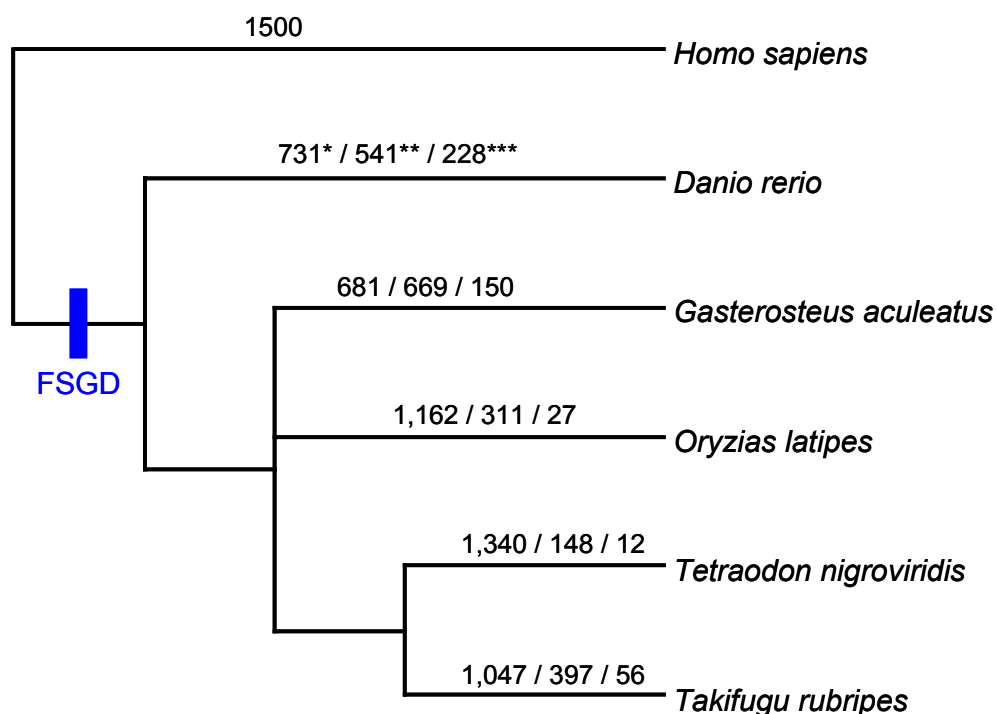


Fig. 2. Differential retention and loss of duplicated genes during teleost diversification. The topology is adopted from (Negrisolo et al. 2010). \*: retention of both copies; \*\*: retention of one copy; \*\*\*: loss of both copies.

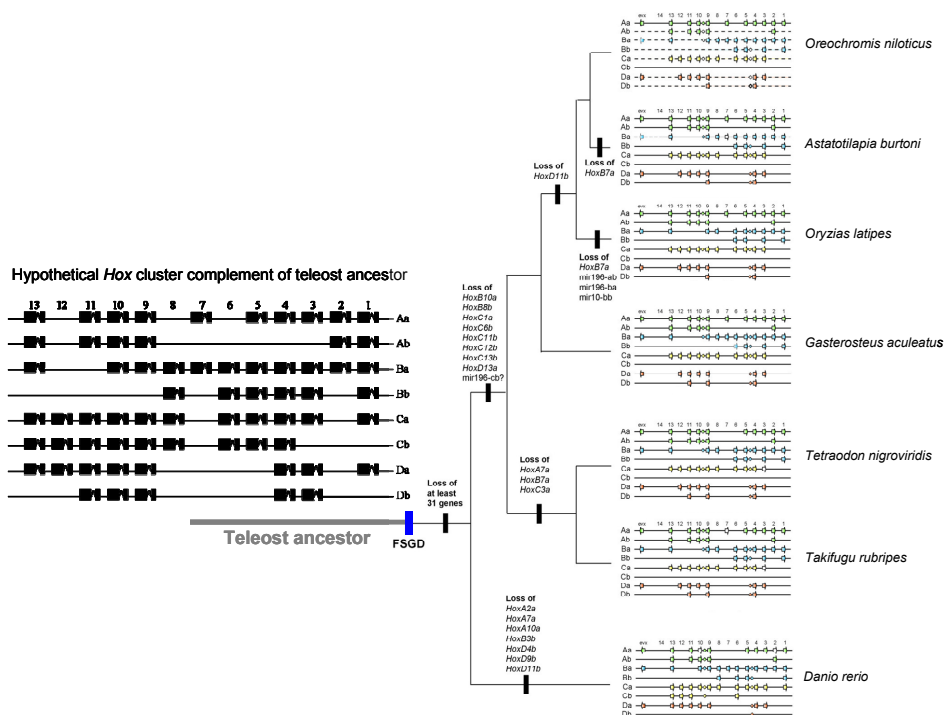


Fig. 3. *Hox* gene clusters, the best-studied examples of differential duplicate gene retention and loss in teleosts. Hypothetical *Hox* clusters of the teleost ancestor (modified from Guo et al. 2010), and *Hox* clusters of teleost model fish species, together with specific gene loss events shown on a phylogenetic tree of select fish species (adapted from Hoegg et al. 2007).

was lost independently in the Otocephala and Euteleostei after the FSGD event. The ongoing process of *Hox* gene loss and retention in teleosts illustrates again that degeneration of functionally important duplicated genes can last for hundreds of millions of years after the FSGD event.

#### 4. Molecular evolution of duplicated genes

We next wished to study patterns of sequence evolution in the 1,500 duplicate gene families we had identified. To this end, we downloaded both nucleic acid and amino acid sequences for genes in these families. For each species, we retained only one gene copy in each duplicated clade (Figure 1B) for further analysis, and discarded all other copies in those gene families where additional duplications have occurred after the FSGD event. We then aligned the amino acid sequences within each gene family with MUSCLE (Edgar



2004), and calculated DNA alignments from protein alignments with RevTrans (Wernersson and Pedersen 2003). The following computations were then done on the new DNA alignments. We estimated the nucleic acid evolutionary distance between fish genes and their human orthologs using the LogDet nucleotide substitution model (Tamura and Kumar 2002) in PHYLIP-3.6b (Felsenstein 2004).

Previous studies show that duplicated genes in yeast often diverge asymmetrically (Kellis et al. 2004), meaning that one copy evolves significantly faster than the other. We asked whether this is also the case for teleost duplicates. To this end, we compared evolutionary distances of duplicated genes with their human orthologs within the 1,500 gene families we had identified. There is indeed evidence for asymmetric evolution between duplicated gene pairs from the FSGD event (Table 2). Average evolutionary distances to the human homologue between members of duplicated gene pairs are significantly different for each of our five teleost species (paired *t*-test:  $P < 4.8 \times 10^{-95}$ ). As all duplicated gene pairs stemming from the FSGD diverged at the same time from their human orthologs, we can directly convert differences between evolutionary distances into differences between evolutionary rates. Taken together, our observations suggest that duplicate genes tend not to accumulate sequence change at the same rate. Our results are consistent with previous works in teleosts (Brunet et al. 2006; Steinke et al. 2006) and yeast (Kellis et al. 2004), and confirm that asymmetric sequence evolution between duplicated genes is a frequent pattern of duplicated gene evolution after a genome duplication event.

	<i>D. rerio</i>	<i>G. aculeatus</i>	<i>O. latipes</i>	<i>T. rubripes</i>	<i>Te. nigroviridis</i>
Duplicate_L	0.613 ± 0.243	0.607 ± 0.229	0.621 ± 0.230	0.623 ± 0.229	0.614 ± 0.224
Duplicate_S	0.529 ± 0.213	0.526 ± 0.200	0.536 ± 0.195	0.535 ± 0.195	0.505 ± 0.182
P-value*	$4.1 \times 10^{-105}$	$4.8 \times 10^{-95}$	$1.9 \times 10^{-165}$	$7.3 \times 10^{-175}$	$8.9 \times 10^{-133}$

Duplicate\_L: duplicated gene in each duplicate pair that has the larger distance to the human orthologue (distances averaged over all duplicate gene families); Duplicate\_S: duplicated gene in each duplicate pair that has the smaller distance to the human orthologue (distances averaged over all duplicate gene families). All means are ± one standard deviation.

\* paired *t*-test

Table 2. Average evolutionary distances of duplicated genes in five teleost species to their human orthologs.

## 5. Conclusion

In summary, we used a phylogenetic method to identify 1,500 duplicated gene families in five teleost species that are likely to have resulted from the FSGD event. Only a small fraction of genes in extant teleost genomes have been retained in the FSGD event. Differential retention and loss of duplicated gene is pervasive in the five species we studied, as is illustrated by genes in the teleost *Hox* gene clusters. Sequence analysis suggests that some duplicated genes pairs may evolve asymmetrically. Our work provides a framework for future studies of the evolutionary trajectory of duplicated genes in the teleost genome.

## 6. Acknowledgement

Support was provided by the National Natural Science Foundation of China (NSFC) to Shunping He.

## 7. References

- Amores A, Force A, Yan YL, et al. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* 282:1711-1714.
- Bridges CB. 1936. The bar "gene" a duplication. *Science* 83:210-211.
- Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* 23:1808-1816.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540-552.
- Chambers KE, McDaniell R, Raincrow JD, Deshmukh M, Stadler PF, Chiu CH. 2009. Hox cluster duplication in the basal teleost *Hiodon alosoides* (Osteoglossomorpha). *Theory Biosci* 128:109-120.
- Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, Venkatesh B. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol* 21:1146-1151.
- Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perriere G. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 21: 2596-2603.
- Duret L, Mouchiroud D, Gouy M. 1994. HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res* 22:2360-2365.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792-1797.
- Felsenstein J. 2004. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
- Guo B, Gan X, He S. 2010. Hox genes of the Japanese eel *Anguilla japonica* and Hox cluster evolution in teleosts. *J Exp Zool B Mol Dev Evol* 314:135-147.
- Hoegg S, Boore JL, Kuehl JV, Meyer A. 2007. Comparative phylogenomic analyses of teleost fish Hox gene clusters: lessons from the cichlid fish *Astatotilapia burtoni*. *BMC Genomics* 8:317.
- Hoegg S, Brinkmann H, Taylor JS, Meyer A. 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J Mol Evol* 59:190-203.
- Jaillon O, Aury JM, Brunet F, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431:946-957.
- Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275-282.

- Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA. 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res* 19:1404-1418.
- Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428:617-624.
- Kurosawa G, Takamatsu N, Takahashi M, et al. 2006. Organization and structure of hox gene loci in medaka genome and comparison with those of pufferfish and zebrafish genomes. *Gene* 370:75-82.
- Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151-1155.
- McGinnis W, Krumlauf R. 1992. Homeobox genes and axial patterning. *Cell* 68:283-302.
- Meyer A, Van de Peer Y. 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays* 27:937-945.
- Negrisol E, Kuhl H, Forcato C, Vitulo N, Reinhardt R, Patarnello T, Bargelloni L. 2010. Different Phylogenomic Approaches to Resolve the Evolutionary Relationships among Model Fish Species. *Mol Biol Evol* 27:2757-2774.
- Ohno S. 1970. Evolution by gene duplication. Springer-Verlag, New York.
- Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, Perriere G. 2009. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* 10 Suppl 6:S3.
- Prohaska SJ, Stadler PF. 2004. The duplication of the Hox gene clusters in teleost fishes. *Theory Biosci* 123:89-110.
- Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440: 341-345.
- Semon M, Wolfe KH. 2007. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet* 23:108-112.
- Stephens SG. 1951. Possible Significance of Duplication in Evolution. In: M Demerec, editor. *Advances in Genetics*: Academic Press. p. 247-265.
- Studer R, Duret L, Penel S, Robinson-Rechavi M. 2008. Pervasive positive selection on duplicated and non duplicated vertebrate protein coding genes. *Genome Res* 18:1393-1402.
- Sugino RP, Innan H. 2005. Estimating the time to the whole-genome duplication and the duration of concerted evolution via gene conversion in yeast. *Genetics* 171: 63-69.
- Tamura K, Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol Biol Evol* 19:1727-1736.
- Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res* 13:382-390.
- Taylor JS, Van de Peer Y, Braasch I, Meyer A. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos Trans R Soc Lond B Biol Sci* 356:1661-1679.
- Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably

between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci U S A* 101: 1638-1643.

Wernersson R, Pedersen AG. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res* 31:3537-3539.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708-713.

# Detection and Analysis of Functional Specialization in Duplicated Genes

Owen Z. Woody and Brendan J. McConkey

*University of Waterloo*

*Canada*

## 1. Introduction

Gene duplication has long been recognized as a powerful mechanism facilitating development and evolution in genomes. Duplication events produce additional copies of genomic information, perhaps including one or more genes. While in some cases these duplicated elements may be of immediate benefit (i.e. increasing availability and effective dosage of a desired gene product), often they are initially at least somewhat redundant, and either neutral or mildly detrimental to the fitness of the organism. It is perhaps no surprise, then, that the majority of duplicated genes are quickly deactivated by mutations abolishing transcription or translation. Some duplicated genes, however, survive and persist, suggesting that their retention has some benefit. Many of these genes seem to have acquired properties that distinguish them from their progenitors – they may be expressed in a novel tissue type, for example, or differ in their functional specificity. In these cases, it appears as though duplication has facilitated evolution, either by allowing specialization and refinement or, perhaps most intriguingly, generating genes free to mutate and acquire ‘novel’ functions. These retained duplicates form a family of genes related through common ancestry. As a result of their common origin, gene sequences within gene families are often quite similar, complicating the task of assigning them unique and specific functions. As such, there has been a significant effort to study and characterize the evolution of function in the aftermath of a duplication event. This chapter will briefly cover the various modes of gene duplication, and then will focus on the various functional outcomes of duplication. The theoretical models for functional specialization following a duplication event are discussed, as are practical techniques for applying these models to observed gene duplicates.

## 2. Mechanisms of duplication

There are several different mutational mechanisms through which gene duplicates can be produced. Depending on the type of event, the nature and scale of what is duplicated can differ significantly; single genes may be copied, with or without their peripheral regulatory elements, or entire genome can be duplicated. While each mechanism ultimately results in the duplication of one or more genes, the mechanisms differ in three key respects; how much regulatory information the duplicated genes retain, where the duplicates are integrated into the genome, and how many interaction partners are duplicated. Duplication

mechanisms can be broadly categorized into three groups – DNA/RNA-mediated transposition, unequal recombination, and genome doubling/hybridization. These mechanisms all produce paralogs -- homologous genes that are both present in and native to the same genome (in contrast to orthologs, where speciation acts as a 'duplication event' and the homologous genes are components of different genomes). Figure 1 provides a diagram depicting various modes of duplication.

### 2.1 DNA/RNA transposition

DNA/RNA transposition refer to mechanisms by which a specific short nucleotide sequence, either mRNA (as in retrotransposition) or DNA (e.g., transposon-mediated duplication) is copied from one location in the genome to another. The insertion location is essentially random; any compatible destination locus will do, and thus the produced duplicate need not necessarily be located near its progenitor template. RNA-mediated retrotransposition is unique in that it uses post-transcription sequence as a template for the nascent duplicate. Hence, upstream and downstream regulatory sequences lying outside the transcribed gene sequence are not preserved, and the newly produced gene will have most or all introns (and possibly some exons) spliced out. The new gene may also possess a genetically encoded poly-A tail. Since RNA-mediated retrotransposition does not preserve most non-coding elements, the duplicate gene must depend on the a-priori availability or acquisition of promoter/regulatory sequences in order to be transcribed. Absence of these elements effectively means the new gene duplicate is a pseudogene.

DNA-mediated duplications, as mediated by transposons, for example, often retain regulatory information and intron/exon structure. Nonetheless, they still operate on a very specific subsequence of DNA, and elements relocated by DNA-mediated transposition can be inserted in any eligible location in the genome.

### 2.2 Segmental duplication/unequal cross-over

Errors during homologous recombination can produce serial duplications of genetic sequence. Unequal crossing-over is an error stemming from the mis-alignment of homologous chromosomes during mitosis/meiosis. Ordinarily, homologous sequences are aligned and cross-over events result in balanced exchanges of sequence information across chromosomes. An abundance of repetitive sequences can, however, cause chromosomes to misalign, in which case a segment of one chromosome is inserted into its sister chromatid (thus producing a duplication and a reciprocal deletion).

Since multiple rounds of unequal crossing over tend to gradually inflate the number of candidate repeat regions, some genomic regions are hotbeds for sequence duplication and can give rise to a large number of duplicate genes in series. These serially arranged duplicates are referred to as “tandem duplicates”. These tandem gene arrays are highly localized in the genome, and tandem duplicates retain most or all of their intron/exon structure and peripheral non-coding elements. Unequal crossing-over also plays a role in the generation of copy number variations (Redon et al., 2006).

### 2.3 Whole genome duplication/allopolyploids

In some circumstances, errors during segregation can produce diploid gametes, and the fusion of these diploid gametes can result in a complete doubling of genomic content (all chromosomes present in duplicate). While very rare, these whole genome duplication

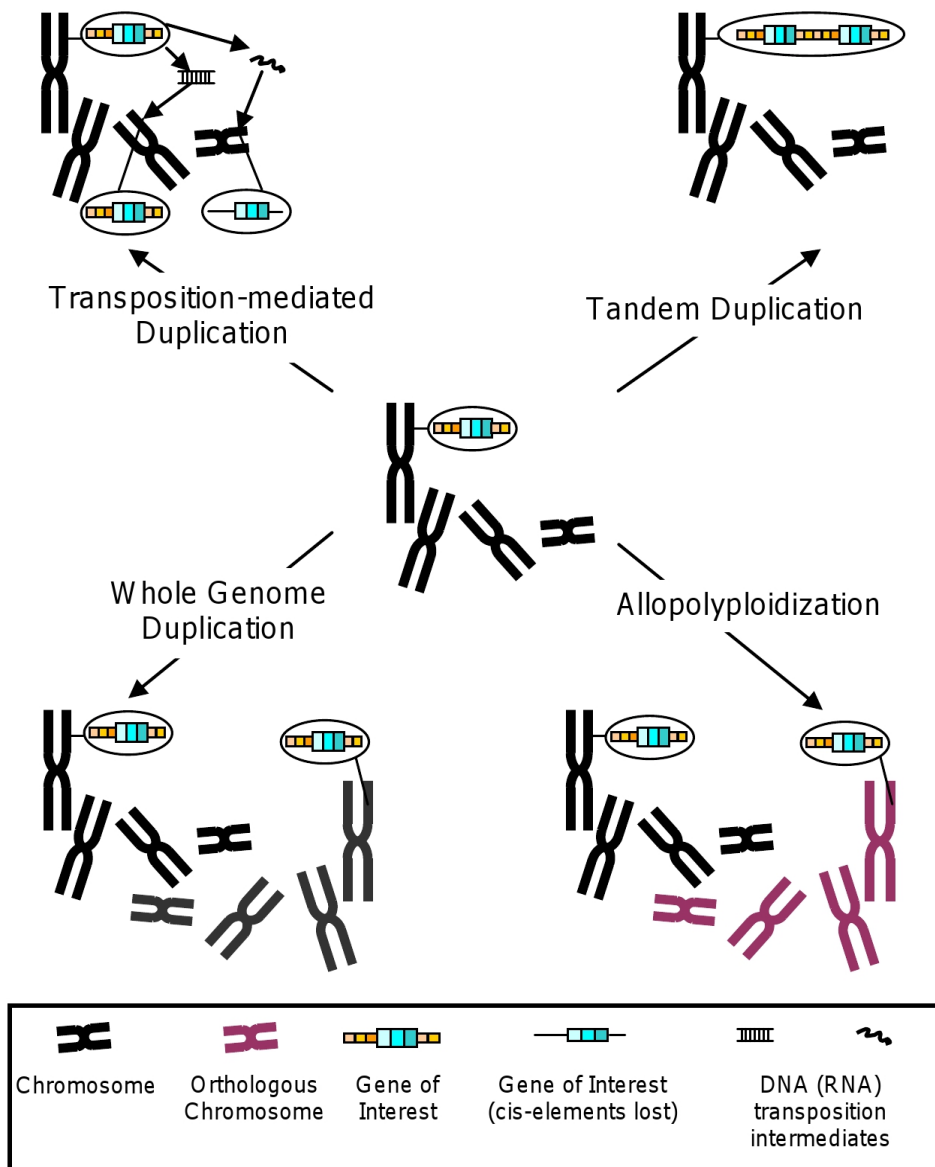


Fig. 1. Modes of gene duplication. Upper left: transposition mediated by either a DNA or RNA intermediate can produce gene locations at distant locations in the genome. RNA intermediates retain little of the regulatory sequence surrounding the parent gene. Upper right: Errors during homologous recombination can produce tandem arrays of genes, situated in series. Bottom Left: Doubling of all chromosomes will produce duplicates of all genes in the genome. Bottom Right: Allopolyploids contain genomes from two compatible species, with duplicate gene pairs from formerly orthologous genes.

(WGD) events have a dramatic impact on the content of the genome, and a number of WGD events have been hypothesized in the history of various lineages (Van de Peer et al., 2009). By their nature, WGD events result in the duplication of all loci, preserving non-coding elements, intron/exon structure, and even overall stoichiometry within gene/protein interaction networks. Interestingly, it has been observed that lineages that undergo separate, distinct WGD events (in this case, *Xenopus tropicalis* and zebrafish) often ultimately retain similar (i.e. orthologous) duplicates – that is to say, WGD duplicates that becomes fixed in one lineage, were also often fixed in the other (Semon & Wolfe, 2008). Pairs of duplicate genes that arose through WGD are sometimes referred to as Ohnologs (Turunen et al., 2009). WGD events are relatively common in plant lineages, which may have interesting implications for the evolution of gene regulation (Lockton & Gaut, 2005).

Allopolyploids are a variant of whole genome duplications in which the diploid gametes come from two different species. These genomic hybrids contain two formerly independent complete genomes. The most commonly studied allopolyploids are plants, though a number of examples have been documented elsewhere in the animal kingdom (including the model organism *Xenopus Laevis*). Duplicates produced through allopolyploidy (i.e. formerly orthologous genes now present in the same organism) are often referred to as “homeologs” (Flagel et al., 2008).

### 3. Defining gene function

One significant hurdle to the discussion of duplicate functional specialization is defining gene function. Gene function may be broken into two broad categories – regulation and gene product (MacCarthy & Bergman, 2007). Regulation encompasses the “when, where, why, and how much” aspects of a gene’s transcription – non-coding elements around a gene (such as enhancers and signaling sequences) can direct when a gene should be expressed, and in what quantity. These non-coding elements are responsive to various cellular and environmental triggers. Changes to regulation alone may be sufficient to bring about the specialization of a new duplicate.

Gene products, on the other hand, primarily dictate the “how” in a gene’s function (along with some regulatory and subcellular localization information present in the 5' and 3' untranslated regions (UTRs)). Studies of duplicated genes have focused on changes to various coding sequence properties, such as binding sites, eligible cofactors, indels, and catalytic residues (Turunen et al., 2009). It’s theoretically possible for a duplicate gene to become functionally specialized without any change to its regulation (Des Marais & Rausher, 2008).

It is worth noting that changes falling into these two categories can occur serially or in concert. For example, a change in tissue localization may precede structural mutations adapting a protein to a new environment.

### 4. Theoretical models for duplicate retention and functional specialization

A number of theoretical models have been proposed to describe how a parental gene’s functions can be partitioned between offspring, and how this partitioning affects the chances of these genes to avoid pseudogenization and eventual deletion.

Three archetypal outcomes – specifically, nonfunctionalization, subfunctionalization, and neofunctionalization, are based on concepts typically attributed to Ohno (1970).



Nonfunctionalization describes the situation where one duplicate's expression is abolished, making it invisible to natural selection and thus free to accumulate mutations. While it is technically possible for a nonfunctionalized gene to have its function restored, the vast majority become relics progressively crippled by the accumulation of disabling and deleterious mutations. There has been some interest in studying the impact losing a duplicate via nonfunctionalization has on sibling genes – for example, whole genome duplication events can lead to cases of “ohnologs gone missing”, where a WGD duplicate has been lost (Canestro et al., 2009). Reciprocal duplicate loss has been hypothesized as one means of speciation. Figure 2 depicts two hypothetical duplications and their respective functional specializations.

#### 4.1 Neofunctionalization

Neofunctionalization refers to the scenario where one duplicate gene acquires mutations that allow it to acquire previously unexplored functions, either through changes in regulation (e.g. tissue localization) or coding sequence. Claims of neofunctionalization tend to focus on the generation of new functions, though it should be noted that these developments may also result in the loss of ancestral function(s) (Turunen et al., 2009).

A specific example of neofunctionalization can be found in a recent study of the MADS-box gene family in angiosperms. MADS-box genes are well-known for their role in developmental processes, but the functions of some gene family members have been difficult to determine. Viaene et al. (2010) provide evidence that a group of these genes, the AGL6 subfamily, can be neatly divided into two groups based on duplication history. One of these groups retains the ancestral function of guiding reproductive development, while the other seems to have acquired a novel role in regulating the growth of vegetative tissues.

A second example, describing the functional differentiation of two paralogs in maize, shows that ancestral functions can still be retained even when one duplicate acquires a novel function (Goettel & Messing, 2010). Two paralogs, named p1 and p2, both drive the synthesis of maysin, which in turn contributes to resistance against earworm. In addition, the p1 gene also has a secondary role in controlling the accumulation of red pigments. The authors propose a series of recombination events that describe how these genes acquired their distinct characters.

#### 4.2 Subfunctionalization

Subfunctionalization involves each gene taking upon a complementary subset of the parental gene's functions, such that neither is independently capable of fulfilling all the parental gene's roles. Subfunctionalization is conceptually synonymous with the Duplication, Degeneration, and Complementation (DDC) model. Regulatory subfunctionalization could result in non-overlapping tissue distributions for the nascent duplicates, with the union of the expression profiles matching the parental gene's range.

Jarinova et al. (2008) describe an instance of subfunctionalization the Hox genes of zebrafish. Through a careful analysis of peripheral non-coding elements, the authors show how the two *hoxb* complexes in zebrafish, *hoxb5a* and *hoxb5b*, acquired non-overlapping expression profiles. In particular, the experimental removal of one regulatory element unique to *hoxb5a* resulted in the two paralogs (re)acquiring a similar expression profile.

The idea of structural subfunctionalization is perhaps best captured in the "Escape from Adaptive Conflict" (EAC) hypothesis. Consider a hypothetical gene product with multiple

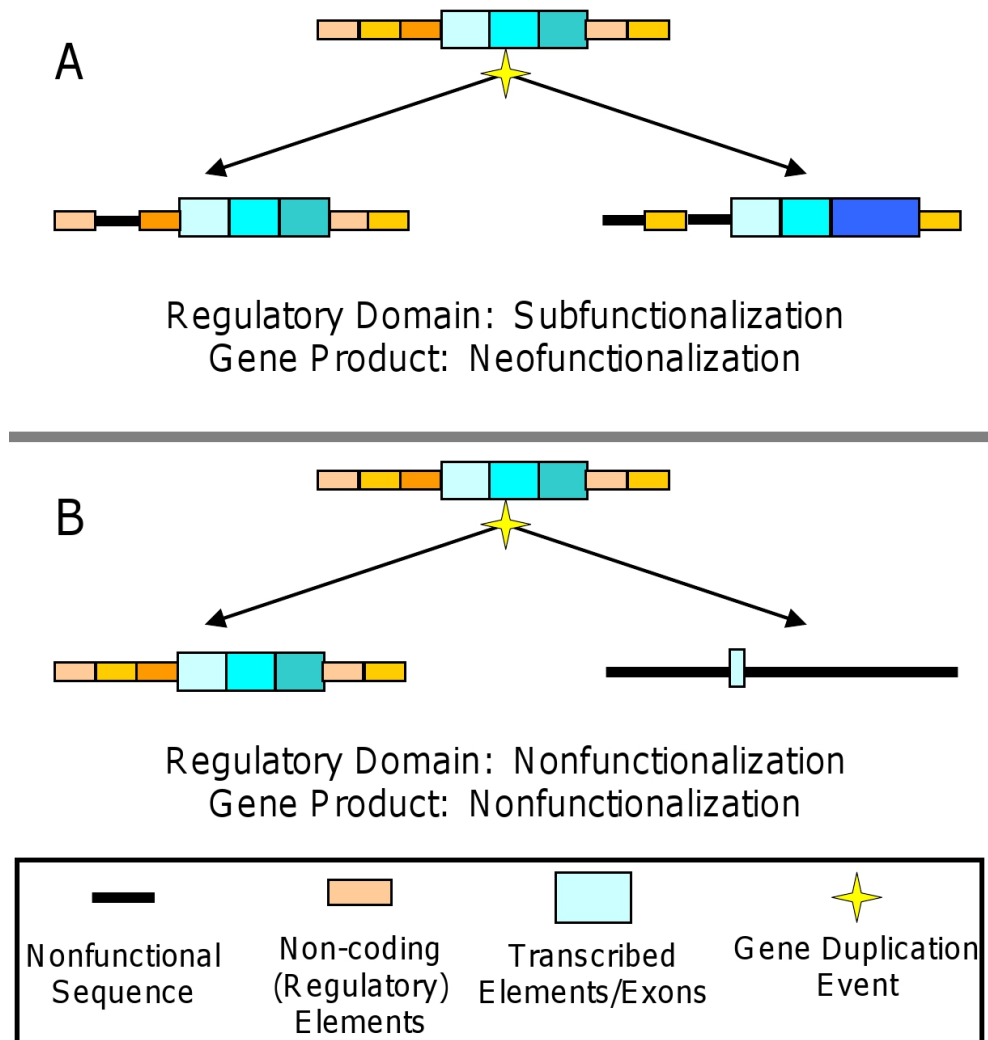


Fig. 2. Possible functional specializations following duplication. Two hypothetical examples showing how retention models can apply either to regulatory regions or gene products. A) Duplicated genes subfunctionalize at the regulatory level, partitioning their parental regulatory domains and suggesting subdivided roles. The gene product, however, has acquired a novel element (i.e. new exon), suggesting neofunctionalization at the coding sequence level. B) Following duplication, one gene loses its regulatory domains and is interrupted by an early stop codon, reflecting nonfunctionalization both at the regulatory and gene product levels.

interaction partners (e.g. an enzyme with two possible substrates). Selection for bifunctionality in this enzyme may limit the binding/catalytic efficiency of either specific reaction -- mutations that improve one may inhibit the other, hence the "adaptive conflict".

Should this gene be duplicated, however, each offspring gene could be free to acquire mutations that optimize binding to one specific substrate, thus escaping the conflict without a loss of functionality. The EAC model essentially describes this process, where a single enzyme with multiple interaction partners gives rise to duplicate genes with more specific but enhanced functionality.

EAC is interesting in that it lies somewhere on the boundary between subfunctionalization and neofunctionalization. Three claims are required to invoke the model: that i) both duplicates accumulate adaptive changes post mutation, that ii) these mutations enhance ancestral functions, and lastly that iii) the ancestral gene was constrained from improving functions (Barkman & Zhang, 2009). The key difference (and challenge) lies in proving the ancestral form was bi-functional. Studies demonstrating the EAC model in action are still relatively uncommon. An early attempt to apply the model to the genes from the anthocyanin biosynthetic pathway of morning glories has come under criticism for not clearly providing these three veins of supporting evidence (Barkman & Zhang, 2009; Des Marais & Rausher, 2008).

The EAC process has also been invoked to describe the evolution of a novel anti-freeze protein in an Antarctic zoarcid fish (Deng et al., 2010). The authors demonstrate that an ancestral gene had a rudimentary ice-binding affinity in addition to its primary catalytic function (a sialic acid synthase), and that a duplication event allowed one copy of this gene to abolish this ancestral function and refine its ice-binding capability. The discussion includes a careful analysis of the three EAC criteria listed above.

While duplicated genes are generally relegated to one of the fates listed above, a number of case studies have shown that recent duplicates can maintain identical functional profiles. One possible explanation for this is that the duplicates have acquired mutations that have restored the “status quo” that was present prior to duplication. If mutations cause the sum of the duplicate genes' expression levels to equal the expression level of their ancestor, both genes could experience some level of selective pressure to maintain expression despite being fully redundant. Ganko et al. (2007) observe that a vast majority of duplicates, regardless of duplication mechanism, showed asymmetric expression, with one gene consistently showing higher levels of expression than its sibling across all tissues. This suggests that a limited form of subfunctionalization may play an initial role in the retention of duplicates. Asymmetrical expression divergence was also observed in a study of duplicated genes in the fly, with a tendency for the “parent” gene of the duplicate to have high expression levels (Langille & Clark, 2007). Interestingly, Qian et al. (2010) point out that many gene duplicates are synthetically lethal or deleterious, and they suggest that expression load may be shared by both genes after duplication.

#### **4.3 Alternative models and odd cases**

A number of other subtle variations have been proposed to augment these three primary fates. Subneofunctionalization, for example, is a model that argues that subfunctionalization followed by neofunctionalization is a common and sequential process. Subfunctionalization permits a relaxation of selection on various subregions of the gene, which in turn allows the (eventual) evolution of novel functions, suggesting that subfunctionalization is more of a midstep than an endpoint (Johnson & Thomas, 2007).

In addition, while subfunctionalization does not make any *a priori* claims about the proportion of functions lost by each duplicate, it appears that in some cases the losses are

highly asymmetrical. Panchin et al. (2010) show that in human duplicate genes one duplicate appears to remain totally unchanged, while its sibling accumulates the majority of functional (in this case, amino acid) mutations.

Contrary to expectations, many gene duplicate pairs appear to be retained despite total apparent functional redundancy. A relatively recent model has been proposed to explain this phenomenon. The theory, coined "originalization", uses arguments based on purifying selection and recombination to support the preservation of both duplicate copies (i.e. prevent non-functionalization) for an extended period of time (Xue & Fu, 2009; Xue et al., 2010).

It has also been suggested that models of duplicate retention are focusing on too small a unit, and that protein interaction networks (themselves composed of a number of co-expressed and functionally related proteins) provide a more coherent perspective on the size of perturbation required to have a phenotypically relevant effect (MacCarthy & Bergman, 2007). The authors argue that cases of regulatory subfunctionalization and neofunctionalization often have no phenotypic consequence on the output of a protein network, and are thus effectively neutral for longer than duplicate-oriented models would suggest. A number of studies have reported an unexplained level and duration of retention for redundant duplicates (Skamnioti et al., 2008).

The relative importance of these models to the retention of duplicates is a subject of continued interest. Whole genome duplication events, which effectively introduce a paralogous copy of every gene in the genome, present an opportunity to tally the cases for which each model applies best. For a review of the relative importance of these various retention models specifically as they pertain to duplicates produced in plant WGD events, see (Edger & Pires, 2009).

## 5. Factors affecting the rate and trajectory of duplicate gene divergence

### 5.1 Gene conversion

Gene conversion describes the process by which the sequence content of one genetic locus is used as a template to alter and "paste over" the genetic sequence at a distal location. Gene conversion has the potential to enforce similarity across duplicate loci, both in terms of regulation and structure. A recent study on duplicated segments in a pair of *Drosophila* species made noted several anomalies that were suggestive of gene conversion. Interestingly, the edges of duplicated regions accumulated distinguishing mutations faster than more central regions, suggesting that these regions were being maintained by gene conversion and that the size of the region being converted was gradually being reduced by sequence mutations near the borders. Furthermore, paralogs near the boundaries of duplicated segments showed more divergence than those located near the centre (Osada & Innan, 2008). The authors note that this phenomenon could result in misleading estimates of synonymous divergence, as the conversion process would periodically homogenize the two sequences.

The requirements for a neofunctionalized gene to escape gene conversion and achieve fixation have been studied from a population genetics perspective (Teshima & Innan, 2008). The fit of the model is tested on a pair of human opsins, which differ in their light sensitivity.

Additional evidence that gene conversion may play a role in duplicate divergence was found in a study of WGD duplicates in rice. Duplicates that contained subsequences of

particularly high sequence similarity also showed a tendency towards retaining similar regulation profiles. The authors suggest that this was a consequence of the promoter regions being propagated via gene conversion activity acting on the locally similar subsequences (Yim et al., 2009).

## 5.2 Alternative splicing

In general, multiple splice forms (and the potential for these splice variants to have distinct functions) have not received much attention in studies of gene duplication and functional divergence. In a first step towards addressing this oversight, Zhan et al. (2011) studied the potential for alternative splicing in *Drosophila* duplicates. New genes tended to show lower levels of alternative splicing, and the subset of duplicates that retained the potential for multiple spliceforms were expressed in fewer tissues, at lower levels, and had had their expression breadth shifted towards preferential expression in testes. The authors also noted that a duplicate's alternative splicing potential depended on duplication mode, with retrotransposed genes being copied with a specific and frozen configuration of exons/introns.

## 5.3 Properties of genes that influence duplicate specialization

The rate at which duplicated genes acquire novel functions is of great interest. Studies have been done to compare the standard metrics of gene evolution (synonymous distance, non-synonymous distance) to measures of functional differentiation across duplicate genes. While initial studies demonstrated only a weak correlation between expression divergence and sequence divergence, subsequent studies have drawn attention to a number of gene parameters that strongly influence the rate and extent of functional differentiation across duplicates.

The mode of duplication has been cited in multiple instances as an important determinant of eventually retention/functional specialization. In a study comparing the functional evolution of genes duplicated through different mechanisms, Ganko et al. (2007) found that WGD-derived duplicates tended to be expressed at higher levels and were more broadly expressed (in contrast to duplicates derived in smaller scale duplications). Wang et al. (2010) found that tandem gene duplicates tended to have conserved function, whereas retrotransposed genes were more likely to have undergone EAC.

Duplicate genes can differ in their tissue distribution, and certain tissues seem to have a greater propensity to adopt genes with novel function than others. In particular, novel duplicates show a tendency towards expression in the testes. Langille and Clark (2007) found that retrotransposed duplicates in particular showed testis-biased expression. Mikhaylova et al. (2008) also found that duplicated genes expressed in the testes tended to show particularly divergent expression across species. Gallach (2010) illustrate a trend for mitochondrial-associated proteins to preferentially fixate in autosomes (i.e. avoiding the X chromosome), and to have a strong tissue bias towards testes expression.

Han et al. (2009) revealed an interesting trend for duplicate genes that had been relocated (i.e. created by transposition). In instances where these duplicated genes showed asymmetric sequence evolution, in more than 80% of cases it was the relocated gene that showed stronger support for positive selection. This suggests an important role for chromosomal distribution in the evolution of gene function and duplicate divergence. A study by Tsankov et al. (2010) also showed that local chromatin organization (i.e.

nucleosome positions) has a strong effect on gene expression, which suggests that translocated duplicates may show expression divergence by virtue of chromosomal position alone. In addition, Ren et al. (2005) found that tandem duplicates that shared expression domains tended to have dissimilar sequence-based functions. Shoja et al. (2007) noted that tandem gene duplicates tended to show a relationship between expression divergence and chromosomal distance.

In their work on the possible action of gene conversion on the evolution of duplicated segments in *Drosophila*, Osada and Innan (2008) noted that duplications lying near the edges of duplicated segments showed more sequence divergence, suggesting that sublocation within a duplicated segment is an additional factor to consider in studies of duplicate divergence.

The broad functional category to which a gene belongs can also influence its freedom to explore divergent functions. In an analysis of genes in the rice genome produced through a specific WGD, Yim et al. (2009) found that duplicate genes with divergent functions showed a significant enrichment towards metabolism-related activity. Langille and Clark (2007) showed that “cell physiological process” genes were particularly amenable to duplication via transposition. Perhaps reflecting similar functional pressures, Li et al. (2010) found that subcellular localization also influenced the divergence of expression between duplicate genes.

The mode of retention may also depend on the amount of selective pressure acting on its coding sequence. Semon and Wolfe (2008) showed that duplicates undergoing slow rates of sequence evolution seemed particularly prone to regulatory subfunctionalization. This observation is echoed in Arnaiz et al. (2010), who find that duplicate pairs in *Tetraurelia* with divergent expression profiles were unlikely to undergo sequence subfunctionalization. Li et al. (2009) found that the mode of duplication had a substantial effect on the degree of expression divergence between duplicates, based on microarray expression profiles of rice tissues.

Nielsen et al. (2010) suggest that genes under strong selective pressure produce duplicates that are quickly nonfunctionalized, suggesting low tolerance for (poisonous) isoforms of essential products. Thus, a gene’s essentiality and, by consequence, age, may both determine the extent to which gene duplicates may be retained.

## 6. Tools for measuring gene regulation

While a gene’s regulatory control is partly controlled by its complement of non-coding elements (as well as its genomic location, e.g., proximity to histones/heterochromatin), efforts to predict regulation from sequence alone have met with limited success, owing to non-linear interactions between various regulatory domains (Jarinova et al., 2008). A separate study found that transcription factor binding site turnover was insufficient to explain cis-regulatory evolution across orthologs (Venkataram & Fay, 2010).

Since accurate predictions of gene regulation based on genomic context and peripheral regulatory elements remain elusive, most studies of gene regulation depend on empirical measurements of gene products (i.e. mRNA or protein) as evidence for a gene’s expression under given conditions. Tools for quantifying the abundance of specific mRNA and/or protein species, such as PCR and Western blots, are standard laboratory techniques.

Within the past decade, however, a number of high-throughput technologies have become available that allow the localization and abundance of gene products to be measured empirically on a genomic/proteomic scale. At present, the most widely used platform is the microarray, an assay with a very large number of transcript-specific probes. Each probe is specific to a known transcript, allowing the potential for complete coverage of all known and predicted genes in a known genome sequence. Custom arrays can also be built from cDNA libraries when working with non-model organisms. Databases replete with microarray data are now publicly available for data mining, allowing a gene's expression (or lack thereof) to be profiled across tissues, timepoints, and stimuli. This aggregate gene behaviour is referred to as an "expression profile", and can serve as an empirical proxy of overall gene function. As more microarray data becomes available, the quality of this proxy will improve.

Expression measurement technologies measure gene activation directly, and are agnostic to the regulatory inputs/mechanisms that lead to transcription. In some cases, *cis*-regulatory regions can undergo substantial changes/shuffling without having much effect on the ultimate transcription behaviour of a gene -- transcription measurement technologies can help distinguish these cases from those that have actually changed a gene's expression phenotype (Comelli & Gonzalez, 2009).

In addition to general purpose (i.e. gene, exon) microarrays, several arrays have been designed to be maximally sensitive to differences between closely related genes. Microarrays use probes that measure targets by hybridizing to nucleotides directly via base complementation. Studies have previously demonstrated that the nucleotides at the center of the probe have the most influence on binding strength. In order to minimize the potential for cross-hybridization, some researchers have designed microarrays for comparing closely related genes (e.g. homeologs) by using probes that feature a known distinguishing SNP at the central position in a probe (Chaudhary et al., 2009; Flagel & Wendel, 2010; Flagel et al., 2008; Udall et al., 2006). This design should minimize cross-hybridization, though it should be noted that previous studies have found that cross-hybridization is only of concern when target sequences are >90-95% identical (Rajashekar et al., 2007). For duplicate genes that have highly similar sequences, alternative measurement technologies like deep sequencing can be used to obtain unbiased paralog-specific expression profiles.

Quantitative proteomics techniques such as iTRAQ (Burkhart et al., 2011) or 2D differential in-gel electrophoresis provide a similarly high-throughput platform for the quantitation of protein abundance. The data differs from microarray data in two respects -- the identities of quantified proteins are often not known in advance, and the coverage of the proteome is not complete and is sensitive to experimental parameters. However, protein abundances may be a more accurate reflection of gene action, as proteins are the active products of genes in most cases and mRNA abundance doesn't always correlate with protein abundance.

Gibson and Goldberg (2009) conducted a study on yeast duplicates using a novel metric of functional differentiation -- number and type of protein interactions. The authors used both affinity-precipitation mass spectrometry and yeast-2-hybrid assays to construct networks of protein interactions, and then sought to test whether the patterns of functional differentiation better fit models of subfunctionalization or neofunctionalization. Their work expands on previous studies that describe the functional evolution of the genome/proteome in terms of the growth of (novel) protein interactions. They illustrate how existing methods

overlook self-self interactions in the parents/progeny, and propose means of avoiding this bias. In general, they found that subfunctionalization was the prevalent driver of protein interaction network evolution.

Recently, sequencing and mass spectrometry have both achieved levels of throughput that make it possible to survey the transcriptome or proteome directly. While these technologies have considerable promise as a source for expression data, at present there are less data available from these platforms (but see Harhay et al., 2010). However, the essential idea of the expression profile holds constant, irrespective of the specific sort (and indeed, mixture) of data that is mined.

## 7. Models of parental gene function

Since available sequence data is generally restricted to present-day organisms, it is not directly possible to measure a gene's function pre- and post-duplication. As such, when presented with a pair of paralogs, it is not often clear which genes have retained the "ancestral function", if any. In this section, a number of proposed techniques for estimating pre-duplication function are described. These techniques can be broadly broken into two categories – techniques which seek to find an appropriate reference organism elsewhere in nature (i.e. a sister species with a somewhat divergent duplication history), and techniques which attempt to estimate/reconstruct ancestral function from those observed in extant species.

If gene information is available for two closely related species, it is often possible to find a number of examples where a gene duplication event has only occurred in one of the two species. In these cases, paralogs in one genome will correspond to a single ortholog in another. By comparing the functions of paralogs to an unduplicated ortholog, it may be possible to infer which of the two paralogs has undergone more dramatic functional changes. Unfortunately, this approach is restricted to genes present in a 2-to-1 fashion, and even in these cases caution must be taken to ensure the duplication event truly post-dates the speciation event.

One interesting variant of this strategy is to use distantly related members of the same gene family from within the same species (Panchin et al., 2010). Since most genes belong to families with several members, recent duplicates can take advantage of ancient and highly diverged gene family members to serve as a proxy for an orthologous outgroup. This process is useful for calculating rates of sequence evolution between recent duplicates.

Comparisons between lineages which have and have not undergone WGD events can also shed light on the evolution of function post-duplication. For example, Kassahn et al. (2009) used mouse orthologs as reference points for evaluating the post-WGD expression divergence in duplicate genes from five teleost fish species, suggesting that this approach is viable even when the organisms being compared are distant relatives.

Allopolyploids present a unique opportunity for studying gene evolution in the aftermath of widespread duplication. Allopolyploids are hybrids of distinct species, and in many cases the unhybridized lineages have persisted alongside their allopolyploid cousins to the present day. In these cases, the history of gene functional evolution can be inferred by examining gene expression behaviour at four stages: pre-hybridization (the two present day diploid parental strains), day zero hybridization (a cross of the two modern day parentals to



create an "F1" allopolyploid), and post-hybridization (the present day allopolyploid). Thus, the functions of both parental genes can be compared to novel and mature hybrids, revealing the immediate effects upon and eventual trajectory of functional evolution.

The utility of this approach can be seen in Chaudhary et al. (2009), where the functional profiles of homeologous genes could be succinctly depicted as two-component pie charts. The dominance of one genome's homelog over another can be visualized as an unequal partitioning in the pie, and changes to this partitioning following the transition from diploidy to allopolyploidy mark possible instances of functional specialization.

However, in many cases there are no suitable extant orthologs available to serve as models for ancestral gene function. In these cases, there are a number of algorithms for estimating ancestral gene function based directly on the functions of descendent (and other related) genes. Estimation methods can try to infer both gene regulation and gene sequence/structure from present day data.

Microarray-based gene expression profiles have been used in several efforts to estimate ancestral gene function. In a study of stress response genes in *Arabidopsis*, estimates of ancestral gene function were constructed using BayesTraits (Pagel & Meade, 2006), with the present day response profiles used as primary data. For each extant stress response gene, responses to various stresses were coded based on expression level changes (up-regulation, down-regulation, no response). By adjusting the parameters of the Bayestraits program, the authors were able to select a model for gain/loss of response behaviour. This information, when combined with phylogenetic trees mapping out the sequence relationships for each gene family, allowed estimates of the stress response behaviour of ancestral genes (internal nodes on the tree) (Zou et al., 2009).

Another microarray-based approach was explored in Doxey et al. (2007). The study examined the beta-(1,3)-glucanase gene family in *Arabidopsis*, using expression profiles constructed from microarray measurements on tissue and stress response patterns. The expression data for all genes in the family were grouped using hierarchical clustering, such that genes with similar (correlated) expression profiles were grouped together. Based on this clustering, genes were assigned labels according to their functional groups, and these labels were then used as primary data for the reconstruction of ancestral states on the gene family phylogenetic tree via parsimony. Using this approach, the expression profile of ancestral, pre-duplication sequences could be estimated from the values reconstructed on the tree.

This approach of reconstructing gene functions as characters on a gene phylogenetic tree has a lot of potential, as it allows all members of a gene family to contribute information about the functional breadth explored in a gene family. The exact quantity reconstructed on the tree can vary from simple binary tissue presence/absence (Karanth et al., 2009) to the exact expression abundance as measured by a high-throughput assay (Guo et al., 2007; Li et al., 2005; Oakley et al., 2005).

There have also been efforts to reconstruct ancestral gene sequences, with the hope of reconstructing gene function. Working from the extant variety of fluorescent proteins, Field and Matz (2010) modeled the evolution of fluorescence color in the family by estimating and producing gene sequences at the internal nodes of the fluorescent protein family phylogenetic tree. By producing proteins based on the estimated ancestral sequences, the authors were able to estimate the fluorescence colors of evolutionary intermediates in the family.

## 8. Making a case for duplicate specialization

The most important techniques for studying functional specialization focus on different aspects of gene function, but are all generally associated with the task of distinguishing the roles and fates of duplicate genes. Figure 3 provides a diagram summarizing the various aspects of gene function that are amenable to these techniques.

### 8.1 Biochemical function and analysis

Unequivocal evidence for functional specialization can be drawn from studies of enzyme kinetics. By measuring the substrate affinity and catalytic rates of enzymes, for example, it is possible to quantitatively measure differences in performance between duplicate genes. Biochemical approaches are very labor intensive and only readily applicable to certain classes of genes, but the evidence they provide is direct and readily interpreted.

In a study highlighting the potential importance of the EAC model, Des Marais and Rauscher (2008) used enzyme affinity assays to demonstrate the enzymatic function of paralogous anthocyanin biosynthetic pathway genes in morning glories. Enzyme kinetics were compared across different species that differed by a duplication of a specific enzyme, with the unduplicated ortholog acting as a proxy for the ancestral function.

A recent innovative approach to studying gene function used directed (in lab) evolution to try and encourage a derived gene to revert to an hypothesized ancestral function (Bershtein & Tawfik, 2008). The authors studied the rate of 'reversion' and how this rate varied when various degrees of selective pressure (selecting for the ancestral function, the current function, or both) were applied. By studying the transitional states the gene underwent as its function shifted, the authors found evidence that best fit the subfunctionalization model of duplicate gene evolution.

### 8.2 Expression profiling and reconstruction

Expression profiles (mined from expression assays like high-throughput sequencing) can provide immediate evidence of functional differentiation between duplicated genes (based on divergent, non-overlapping expression behaviour). For example, Yasukawa et al. (2010) use RNA in-situ and reporter analysis to determine precise expression localization and timing of duplicates in *Drosophila*. Rajashekar (2007) used microarray expression profiles to analyze the similarity of duplicates in the hydrophobin gene family in a fungus, *Paxillus involutus*.

By augmenting comparisons of gene expression profiles with reconstructed estimates of parental gene expression (via ancestral character estimation projected onto phylogenetic trees, see section 7), it is further possible to estimate how each gene progeny specialized from its parent following duplication. Case examples are the studies by Doxey et al. (2007) and Zou et al. (2009), both mentioned previously. Zou et al. (2009) reconstructed the expression behaviour in stress response genes in *Arabidopsis*, and with the additional information made available in the estimates of ancestral behaviour, the authors were able to infer patterns of subfunctionalization and neofunctionalization leading to the expression behaviour in extant duplicates. Karanth et al. (2009) reconstructed the ancestral tissue expression patterns of fatty acid binding proteins in zebrafish, revealing

an apparent neofunctionalization event followed by subfunctionalization in a subsequent duplication.

### 8.3 Comparing with a non-duplicated ortholog

As discussed earlier, one effective technique for estimating the function of the ancestor of a pair of duplicate genes is to refer to a related species where the locus is unduplicated. In this case, the assumption is that the orthologous gene is behaving in the related genome as the parental gene was behaving prior to the duplication event. This point of reference makes it possible to distinguish between models of duplicate retention, lending to support towards subfunctionalization versus neofunctionalization, for example.

In a study of zebrafish-specific WGD-produced duplicates, Kassahn et al. (2009) use unduplicated mouse orthologs as a reference, despite the considerable distance separating these two organisms. Multiple gene properties were compared between paralogs and their mouse ortholog, including sequence, structure, and expression information. The authors found support for neofunctionalization in a number of duplicates, and that regulatory changes were far more common than changes to gene products.

In a study of human genes, Panchin et al. (2010) chose to use distantly related gene family members as proxies for ancestors of recent paralogs. They demonstrated that, in many cases, the recent duplicates are evolving asymmetrically, with one duplicate accumulating sequence mutations much faster than its sibling.

Semon and Wolfe (2008) conducted a study comparing the fate of WGD duplicates in *X.laevis*, an allopolyploid, to *X. tropicalis*, a related species that did not undergo any WGD. Expression patterns were compared across 11 tissue types, and related losses of tissue breadth to possible subfunctionalization. In addition to this, the authors also compared the fate of duplicated genes produced through two different large-scale duplication mechanisms by comparing *X.laevis* to zebrafish, a species with a well studied WGD that did not stem from allopolyploidy. They find that duplicates retained in the *X.laevis* duplication were also frequently retained in duplicate in zebrafish, suggesting common influences on the duplicability of these gene varieties.

Another example of a well-studied allopolyploid, cotton, has been discussed in previous sections (Chaudhary et al., 2009; Flagel et al., 2008; Flagel & Wendel, 2010). One unique observation made possible in this system is the phenomenon of transgressive segregation, where the expression profiles of homeologous genes eventually evolve to resemble neither of the parental strains, suggesting a unique adaptation to the presence of two essentially complete genomes within a single cell (Flagel & Wendel, 2010).

### 8.4 Comparing gene product properties

While not as easily assayed as gene expression, the transcribed content of genes (i.e. proteins) can also suggest the gain and loss of functions. As a simple example, the rate of protein sequence evolution can be compared between duplicates by comparing their respective rates of synonymous and non-synonymous mutation. While not necessarily illustrative of the nature of the difference, this method can provide evidence for asymmetrical selection, suggesting one duplicate is acquiring amino acid altering mutations faster than the other (Ganko et al., 2007). Working from a list of 15 of the most asymmetrically diverged WGD-derived protein sequences in *S. cerevisiae*, Turunen et al. (2009) noted substantial indels in addition to changes in important catalytic residues and

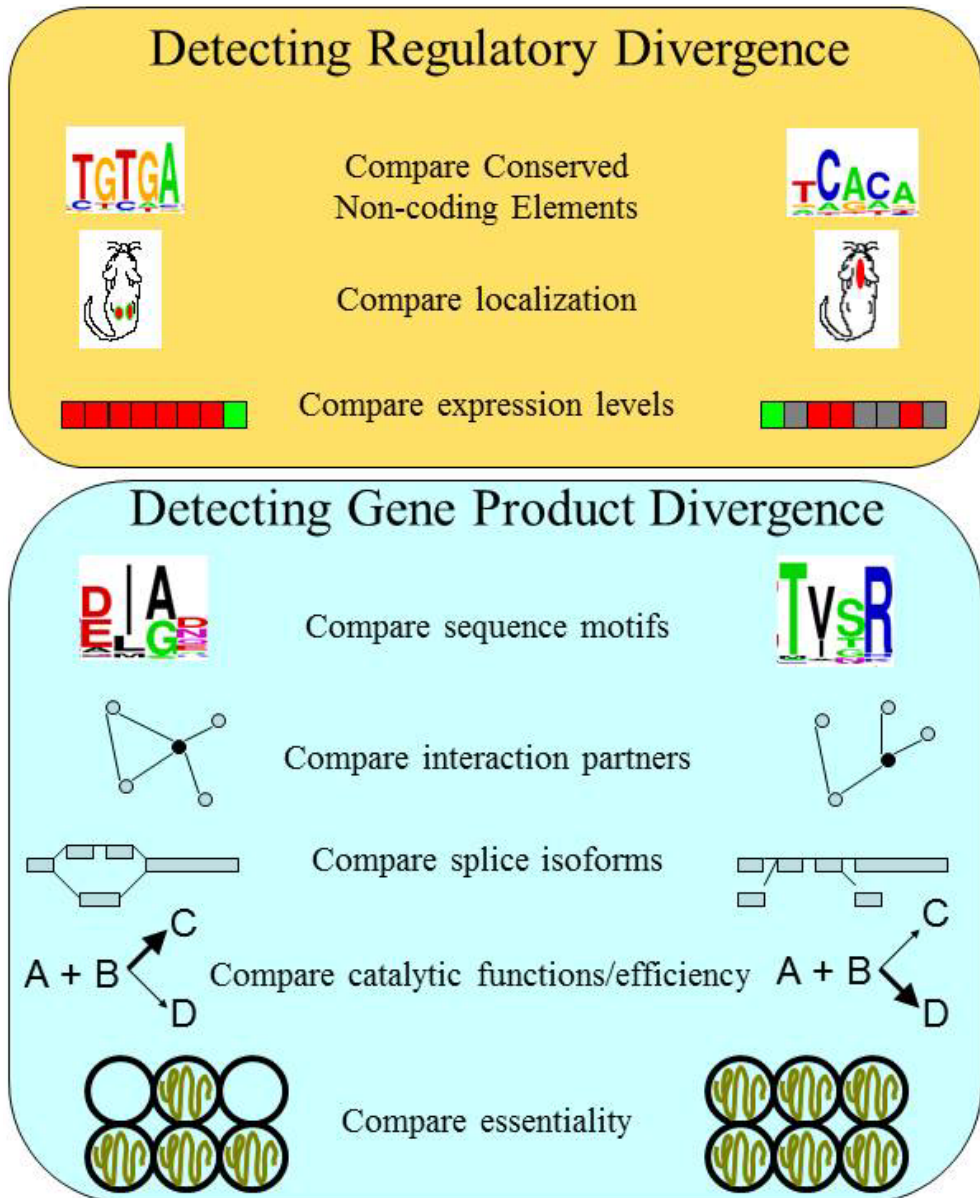


Fig. 3. Gene properties that can be examined for evidence of functional specialization. The top set (orange) are approaches that check for differences in gene regulation; expression levels reflect measurements of transcription in tissues in response to a series of stresses (e.g. as obtained from microarrays). The bottom set (blue) are aspects of the gene product that may differ between duplicates. Sequence logos may be generated using the WebLogo software (Crooks et al., 2004).

active/cofactor binding sites. A literature search seemed to support that many of these highly modified duplicates had acquired novel functions.

Other aspects of gene function, such as the position and number of introns or methylation sites, have also been used to characterize divergence between duplicated genes. For example, Xiong et al. (2010) include intron position in a study of the expansion of the ABC transporter gene family in the ciliate *Tetrahymena thermophila*. In addition to comparing the expression profiles constructed by clustering gene expression data, the authors also compare intron positions to group the family members into functional subcategories (Xiong et al., 2010). A similar study has examined differential splicing forms in duplicate genes of *Drosophila* (Zhan et al., 2011).

Yang et al. (2006) compared the “DNA-binding with one finger” (DOF) gene family across three plant species – rice, *Arabidopsis thaliana*, and poplar. Their multifaceted approach to describing gene function included an analysis of protein motif gain/loss and changes to methylation patterns. Combined with information about gene regulation drawn from microarrays, PCR and massively parallel signature sequencing, the authors compared the relative applicability of various duplicate retention models to the DOF family.

When the information is available, the protein-protein interaction partners of duplicates can also be compared to study duplicate specialization. Nielsen et al. (2010) compared a set of residues in the tail ends of tubulin genes in fruit flies, noting divergence in these regions which may reflect changes in protein-protein interaction partners. Studying the applicability of models like subfunctionalization and neofunctionalization at the level of gene networks has helped integrate duplicate specialization into a broader systems biology context (Gibson & Goldberg, 2009; McCarthy & Bergman, 2007).

## 9. Conclusion

Studies of the evolution of duplicate genes are pushing the field towards more exacting standards and definitions for gene function. Since the rate and extent of duplicate gene specialization is dependent on so many factors, and since novel functions can emerge in so many different ways, integrative approaches will be of paramount importance to understanding this key aspect of genomic evolution. Future studies can benefit in particular from the inclusion of data from gene families as a whole, as this additional information helps both with estimating ancestral gene functions and with evaluating the breadth of function previously covered by related genes.

While empirical evidence of differential catalytic function remains the gold standard for proving functional specialization of duplicated genes, high-throughput studies exploiting the vast quantities of minable expression data provide a cheap and effective means for studying functional specialization at the level of whole gene families. Genomes with annotations beyond expression profiles (such as gene-by-gene interaction profiles and essentiality data) should be helpful for determining the extent to which functional changes at the regulatory level actually impact phenotype.

## 10. References

- Arnaiz, O., Gout, J. F., Betermier, M., Bouhouche, K., Cohen, J., Duret, L., Kapusta, A., Meyer, E. & Sperling, L. (2010). Gene expression in a paleopolyploid: a

- transcriptome resource for the ciliate *Paramecium tetraurelia*. *BMC Genomics*, Vol. 11, No. 547
- Barkman, T. & Zhang, J. (2009). Evidence for escape from adaptive conflict? *Nature*, Vol. 462, No. 7274
- Bershtein, S. & Tawfik, D. S. (2008). Ohno's model revisited: measuring the frequency of potentially adaptive mutations under various mutational drifts. *Mol. Biol. Evol.*, Vol. 25, No. 11, pp. 2311-2318
- Burkhardt, J. M., Vaudel, M., Zahedi, R. P., Martens, L. & Sickmann, A. (2011). iTRAQ protein quantification: A quality-controlled workflow. *Proteomics*, Vol. 11, No. 6, pp. 1125-1134
- Canestro, C., Catchen, J. M., Rodriguez-Mari, A., Yokoi, H. & Postlethwait, J. H. (2009). Consequences of lineage-specific gene loss on functional evolution of surviving paralogs: ALDH1A and retinoic acid signaling in vertebrate genomes. *PLoS Genet.*, Vol. 5, No. 5
- Chaudhary, B., Flagel, L., Stupar, R. M., Udall, J. A., Verma, N., Springer, N. M. & Wendel, J. F. (2009). Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (*Gossypium*). *Genetics*, Vol. 182, No. 2, pp. 503-517
- Comelli, R. N. & Gonzalez, D. H. (2009). Divergent regulatory mechanisms in the response of respiratory chain component genes to carbohydrates suggests a model for gene evolution after duplication. *Plant. Signal. Behav.*, Vol. 4, No. 12, pp. 1179-1181
- Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.*, Vol. 14, No. 6, pp. 1188-1190
- Deng, C., Cheng, C. H., Ye, H., He, X. & Chen, L. (2010). Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proc. Natl Acad. Sci. U.S.A.*, Vol. 107, No. 50, pp. 21593-21598
- Des Marais, D. L. & Rausher, M. D. (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, Vol. 454, No. 7205, pp. 762-765
- Doxey, A. C., Yaish, M. W., Moffatt, B. A., Griffith, M. & McConkey, B. J. (2007). Functional divergence in the *Arabidopsis* beta-1,3-glucanase gene family inferred by phylogenetic reconstruction of expression states. *Mol. Biol. Evol.*, Vol. 24, No. 4, pp. 1045-1055
- Edger, P. P. & Pires, J. C. (2009). Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.*, Vol. 17, No. 5, pp. 699-717
- Field, S. F. & Matz, M. V. (2010). Retracing evolution of red fluorescence in GFP-like proteins from *Faviina* corals. *Mol. Biol. Evol.*, Vol. 27, No. 2, pp. 225-233
- Flagel, L., Udall, J., Nettleton, D. & Wendel, J. (2008). Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. *BMC Biol.*, Vol. 6, No. 16
- Flagel, L. E. & Wendel, J. F. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytol.*, Vol. 186, No. 1, pp. 184-193

- Gallach, M., Chandrasekaran, C. & Betran, E. (2010). Analyses of nuclearly encoded mitochondrial genes suggest gene duplication as a mechanism for resolving intralocus sexually antagonistic conflict in *Drosophila*. *Genome Biol. Evol.*, Vol. 2, pp. 835-850
- Ganko, E. W., Meyers, B. C. & Vision, T. J. (2007). Divergence in expression between duplicated genes in *Arabidopsis*. *Mol. Biol. Evol.*, Vol. 24, No. 10, pp. 2298-2309
- Gibson, T. A. & Goldberg, D. S. (2009). Questioning the ubiquity of neofunctionalization. *PLoS Comput. Biol.*, Vol. 5, No. 1
- Goettel, W. & Messing, J. (2010). Divergence of gene regulation through chromosomal rearrangements. *BMC Genomics*, Vol. 11, No. 678
- Guo, H., Weiss, R. E., Gu, X. & Suchard, M. A. (2007). Time squared: repeated measures on phylogenies. *Mol. Biol. Evol.*, Vol. 24, No. 2, pp. 352-362
- Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C. & Hahn, M. W. (2009). Adaptive evolution of young gene duplicates in mammals. *Genome Res.*, Vol. 19, No. 5, pp. 859-867
- Harhay, G. P., Smith, T. P., Alexander, L. J., Haudenschild, C. D., Keele, J. W., Matukumalli, L. K., Schroeder, S. G., Van Tassell, C. P., Gresham, C. R., Bridges, S. M., Burgess, S. C. & Sonstegard, T. S. (2010). An atlas of bovine gene expression reveals novel distinctive tissue characteristics and evidence for improving genome annotation. *Genome Biol.*, Vol. 11, No. 10
- Jarinova, O., Hatch, G., Poitras, L., Prudhomme, C., Grzyb, M., Aubin, J., Berube-Simard, F. A., Jeannotte, L. & Ekker, M. (2008). Functional resolution of duplicated *hoxb5* genes in teleosts. *Development*, Vol. 135, No. 21, pp. 3543-3553
- Johnson, D. A. & Thomas, M. A. (2007). The monosaccharide transporter gene family in *Arabidopsis* and rice: a history of duplications, adaptive evolution, and functional divergence. *Mol. Biol. Evol.*, Vol. 24, No. 11, pp. 2412-2423
- Karanth, S., Denovan-Wright, E. M., Thisse, C., Thisse, B. & Wright, J. M. (2009). Tandem duplication of the *fabp1b* gene and subsequent divergence of the tissue-specific distribution of *fabp1b.1* and *fabp1b.2* transcripts in zebrafish (*Danio rerio*). *Genome*, Vol. 52, No. 12, pp. 985-992
- Kassahn, K. S., Dang, V. T., Wilkins, S. J., Perkins, A. C. & Ragan, M. A. (2009). Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res.*, Vol. 19, No. 8, pp. 1404-1418
- Langille, M. G. & Clark, D. V. (2007). Parent genes of retrotransposition-generated gene duplicates in *Drosophila melanogaster* have distinct expression profiles. *Genomics*, Vol. 90, No. 3, pp. 334-343
- Li, Q., Liu, X., He, Q., Hu, L., Ling, Y., Wu, Y., Yang, X. & Yu, L. (2011). Systematic analysis of gene expression level with tissue-specificity, function and protein subcellular localization in human transcriptome. *Mol. Biol. Rep.*, Vol. 38, No. 4, pp. 2597-2602.
- Li, Z., Zhang, H., Ge, S., Gu, X., Gao, G. & Luo, J. (2009). Expression pattern divergence of duplicated genes in rice. *BMC Bioinformatics*, Vol. 10 Suppl 6
- Li, Z., Liu, Q., Song, M., Zheng, Y., Nan, P., Cao, Y., Chen, G., Li, Y. & Zhong, Y. (2005). Detecting correlation between sequence and expression divergences

- in a comparative analysis of human serpin genes. *BioSystems*, Vol. 82, No. 3, pp. 226-234
- Lockton, S. & Gaut, B. S. (2005). Plant conserved non-coding sequences and paralogue evolution. *Trends Genet.*, Vol. 21, No. 1, pp. 60-65
- MacCarthy, T. & Bergman, A. (2007). The limits of subfunctionalization. *BMC Evol.Biol.*, Vol. 7, No. 213
- Mikhaylova, L. M., Nguyen, K. & Nurminsky, D. I. (2008). Analysis of the *Drosophila melanogaster* testes transcriptome reveals coordinate regulation of paralogous genes. *Genetics*, Vol. 179, No. 1, pp. 305-315
- Nielsen, M. G., Gadagkar, S. R. & Gutzwiller, L. (2010). Tubulin evolution in insects: gene duplication and subfunctionalization provide specialized isoforms in a functionally constrained gene family. *BMC Evol.Biol.*, Vol. 10, No. 113
- Oakley, T. H., Gu, Z., Abouheif, E., Patel, N. H. & Li, W. H. (2005). Comparative methods for the analysis of gene-expression evolution: an example using yeast functional genomic data. *Mol.Biol.Evol.*, Vol. 22, No. 1, pp. 40-50
- Ohno, S. (1970). *Evolution by Gene Duplication*, Springer-Verlag, 0-04-575015-7, New York
- Osada, N. & Innan, H. (2008). Duplication and gene conversion in the *Drosophila melanogaster* genome. *PLoS Genet.*, Vol. 4, No. 12
- Pagel, M. & Meade, A. (2006). Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. *Am. Nat.*, Vol. 167, No. 6
- Panchin, A. Y., Gelfand, M. S., Ramensky, V. E. & Artamonova, I. I. (2010). Asymmetric and non-uniform evolution of recently duplicated human genes. *Biol. Direct*, Vol. 5
- Qian, W., Liao, B. Y., Chang, A. Y. & Zhang, J. (2010). Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.*, Vol. 26, No. 10, pp. 425-430
- Rajashekar, B., Samson, P., Johansson, T. & Tunlid, A. (2007). Evolution of nucleotide sequences and expression patterns of hydrophobin genes in the ectomycorrhizal fungus *Paxillus involutus*. *New Phytol.*, Vol. 174, No. 2, pp. 399-411
- Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global variation in copy number in the human genome. *Nature*, Vol. 444, No. 7118, pp. 444-454
- Ren, X. Y., Fiers, M. W., Stiekema, W. J. & Nap, J. P. (2005). Local coexpression domains of two to four genes in the genome of *Arabidopsis*. *Plant Physiol.*, Vol. 138, No. 2, pp. 923-934
- Semon, M. & Wolfe, K. H. (2008). Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. *Proc. Natl Acad. Sci. U.S.A.*, Vol. 105, No. 24, pp. 8333-8338
- Shoja, V., Murali, T. M. & Zhang, L. (2007). Expression divergence of tandemly arrayed genes in human and mouse. *Comp. Funct. Genomics*, 60964
- Skamnioti, P., Furlong, R. F. & Gurr, S. J. (2008). The fate of gene duplicates in the genomes of fungal pathogens. *Commun. Integr. Biol.*, Vol. 1, No. 2, pp. 196-198



- Teshima, K. M. & Innan, H. (2008). Neofunctionalization of duplicated genes under the pressure of gene conversion. *Genetics*, Vol. 178, No. 3, pp. 1385-1398
- Tsankov, A. M., Thompson, D. A., Socha, A., Regev, A. & Rando, O. J. (2010). The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol.*, Vol. 8, No. 7
- Turunen, O., Seelke, R. & Macosko, J. (2009). In silico evidence for functional specialization after genome duplication in yeast. *FEMS Yeast Res.*, Vol. 9, No. 1, pp. 16-31
- Udall, J. A., Swanson, J. M., Nettleton, D., Percifield, R. J. & Wendel, J. F. (2006). A novel approach for characterizing expression levels of genes duplicated by polyploidy. *Genetics*, Vol. 173, No. 3, pp. 1823-1827
- Van de Peer, Y., Maere, S. & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.*, Vol. 10, No. 10, pp. 725-732
- Venkataram, S. & Fay, J. C. (2010). Is transcription factor binding site turnover a sufficient explanation for cis-regulatory sequence divergence? *Genome Biol. Evol.*, Vol. 2, pp. 851-858
- Viaene, T., Vekemans, D., Becker, A., Melzer, S. & Geuten, K. (2010). Expression divergence of the AGL6 MADS domain transcription factor lineage after a core eudicot duplication suggests functional diversification. *BMC Plant. Biol.*, Vol. 10, No. 148
- Wang, Z., Dong, X., Ding, G. & Li, Y. (2010). Comparing the retention mechanisms of tandem duplicates and retrogenes in human and mouse genomes. *Genet. Sel. Evol.*, Vol. 42, pp. 24
- Xiong, J., Feng, L., Yuan, D., Fu, C. & Miao, W. (2010). Genome-wide identification and evolution of ATP-binding cassette transporters in the ciliate *Tetrahymena thermophila*: A case of functional divergence in a multigene family. *BMC Evol. Biol.*, Vol. 10, No. 330
- Xue, C., Huang, R., Liu, S. Q. & Fu, Y. X. (2010). Recombination facilitates neofunctionalization of duplicate genes via originalization. *BMC Genet.*, Vol. 11, No. 46
- Xue, C. & Fu, Y. (2009). Preservation of duplicate genes by originalization. *Genetica*, Vol. 136, No. 1, pp. 69-78
- Yang, X., Tuskan, G. A. & Cheng, M. Z. (2006). Divergence of the Dof gene families in poplar, *Arabidopsis*, and rice suggests multiple modes of gene evolution after duplication. *Plant Physiol.*, Vol. 142, No. 3, pp. 820-830
- Yasukawa, J., Tomioka, S., Aigaki, T. & Matsuo, T. (2010). Evolution of expression patterns of two odorant-binding protein genes, Obp57d and Obp57e, in *Drosophila*. *Gene*, Vol. 467, No. 1-2, pp. 25-34
- Yim, W. C., Lee, B. M. & Jang, C. S. (2009). Expression diversity and evolutionary dynamics of rice duplicate genes. *Mol. Genet. Genomics*, Vol. 281, No. 5, pp. 483-493, 1617-4623
- Zhan, Z., Ren, J., Zhang, Y., Zhao, R., Yang, S. & Wang, W. (2011). Evolution of alternative splicing in newly evolved genes of *Drosophila*. *Gene*, Vol. 470, No. 1-2, pp. 1-6

- Zou, C., Lehti-Shiu, M. D., Thomashow, M. & Shiu, S. H. (2009). Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. *PLoS Genet.*, Vol. 5, No. 7, e1000581

# Predicting Tandemly Arrayed Gene Duplicates with WebScipio

Klas Hatje and Martin Kollmar

*Abteilung NMR basierte Strukturbiologie, Max-Planck-Institut für  
Biophysikalische Chemie, Am Fassberg 11, Göttingen  
Germany*

## 1. Introduction

Since the first high-quality eukaryotic genome assemblies became available the large scale analysis of the origin of new genes came into the focus of many studies (Shoja & Zhang, 2006; Zhou et al., 2008). New genes can originate through multiple mechanisms including gene duplication, gene fusion/fission, exon shuffling, retroposition, horizontal gene transfer, and de novo from noncoding sequences (Long et al., 2003). Although initial models proposed that new copies of genes soon become nonfunctional (Nei & Roychoudhury, 1973; Ohno, 1970) it has since been shown for numerous genes that they retain function through creating redundancy, subfunctionalization, and neofunctionalization (Hahn, 2009; Li et al., 2005; Massingham et al., 2001). While de novo origination from noncoding sequence has been shown to play an unexpectedly important role (Zhou et al., 2008) most of the new genes are derived through duplications. Gene duplicates are normally classified into dispersed and tandem duplicates. Tandem duplications of clusters of genes, single genes, groups of exons, or single exons are thought to be formed by unequal crossing-over events, or misaligned homologous recombinational repair (Babushok et al., 2007; Zhang, 2003). A comparative analysis of the human, mouse, and rat genome has shown that about 15 % of all genes represent tandemly arrayed genes (Shoja & Zhang, 2006). A similar number of about 20 % has been found for the fruit fly *Drosophila melanogaster* (Quijano et al., 2008). All these analyses rely on the particular dataset of annotated genes used and the specific methods for defining genes as tandem genes. However, first annotations of genomes are in most cases done by automatic gene prediction programs, nowadays often supported by incorporating additional EST data, and therefore miss many genes, include artificially fused neighbouring genes, and contain mis-predicted exons and introns. Although these errors seem small, in the case of distinguishing tandem gene duplicates from genomic region duplication and *trans*-spliced genes they are essential. In addition, defining tandem genes by a certain number of nucleotides appearing in-between cannot separate tandem gene duplicates from duplications of small genomic regions. Tandemly arrayed gene duplicates are often conserved between species. Examples are the olfactory receptor genes that constitute a very large gene family of several hundred genes per species in vertebrates (Aloni et al., 2006) and the HOX genes (Garcia-Fernandez, 2005; Zhang & Nei, 1996). While algorithms have been developed to reconstruct the history and evolution of tandemly arrayed genes (Bertrand et al., 2008; Elemento et al., 2002) specific programs are not available for the prediction and local reconstruction of these gene arrays.

WebScipio is a web application to reconstruct genes based on a given protein query sequence and a genomic DNA target sequence (Odrionitz et al., 2008). The reconstruction is done with Scipio (Keller et al., 2008), a post-processing script for the output of a BLAT run (Kent, 2002). BLAT is a very fast tool for the alignment of protein or DNA sequences if these sequences are almost identical. However, BLAT is not able to reconstruct intron and exon borders, it does not identify very short exons and very divergent exons, and it is not able to reconstruct genes spread on several pieces of contiguous DNA (contigs), which is very common in low-coverage genome assemblies. Furthermore, BLAT is not able to identify sequencing and assembly errors like additional or missing bases in exon regions or base substitutions leading to in-frame stop codons. Scipio is able to correct all these errors and extend the BLAT output for the missing sequences of short or divergent exons and of exon borders. In addition, Scipio assembles genes spread on several contigs. WebScipio has been developed as a web interface to Scipio so that the user does not have to install scripts and libraries. Moreover, WebScipio offers access to about 2300 genome assembly files of more than 650 sequenced eukaryotes (July 2011), and provides graphical and human-readable analyses of the results.

Here, we present an extension to the WebScipio web application to search for and predict tandemly arrayed gene duplicates for a given query sequence. This extension is not available via the Scipio command-line script. The user can search for gene duplicates in hundreds of species for which reliable annotations are not available yet, because WebScipio provides access to thousands of genome files.

## 2. Implementation

The new algorithm to predict tandemly arrayed gene duplicates is fully integrated into the web application WebScipio to make it usable for the inexperienced user and to visualize the results for immediate analysis. It was implemented in the Ruby programming language (Ruby Programming Language, 2011) using the BioRuby library (Goto et al., 2010) to handle sequences. WebScipio is based on the web framework Ruby on Rails (Ruby on Rails, 2011), which includes the Javascript libraries Prototype (Prototype JavaScript framework: Easy Ajax and DOM manipulation for dynamic web applications, 2011) and Scriptaculous (script.aculo.us - web 2.0 javascript, 2011). To keep the web application responsive, the search algorithm runs in the background with the help of the Ruby on Rails plug-ins Working (purzelrakete's working at master - GitHub, 2011) and Spawn (tra's spawn at master - GitHub, 2011). To store the user session data, the database backend Tokyo Tyrant is used in combination with Tokyo Cabinet (Tokyo Cabinet: a modern implementation of DBM, 2011). The results of the search are presented as SVG pictures (W3C SVG Working Group, 2011) and several human-readable representations, most notably a detailed alignment of protein query, target DNA sequence, and target translation. The raw results can be downloaded as General Feature Format (GFF) files or as YAML files (The Official YAML Web Site, 2011) for future upload and analysis. Specific results are available in various formats for further inspection, like the human-readable log-files, or publication quality figures, like the SVGs.

### 2.1 Search algorithm

The overall workflow of the search algorithm is shown in Fig. 1. The search for tandem gene duplications is based on the exon-intron structure of a gene generated by Scipio. Thus the first step of the algorithm includes a WebScipio run generating a new gene structure or the upload of an existing Scipio result.

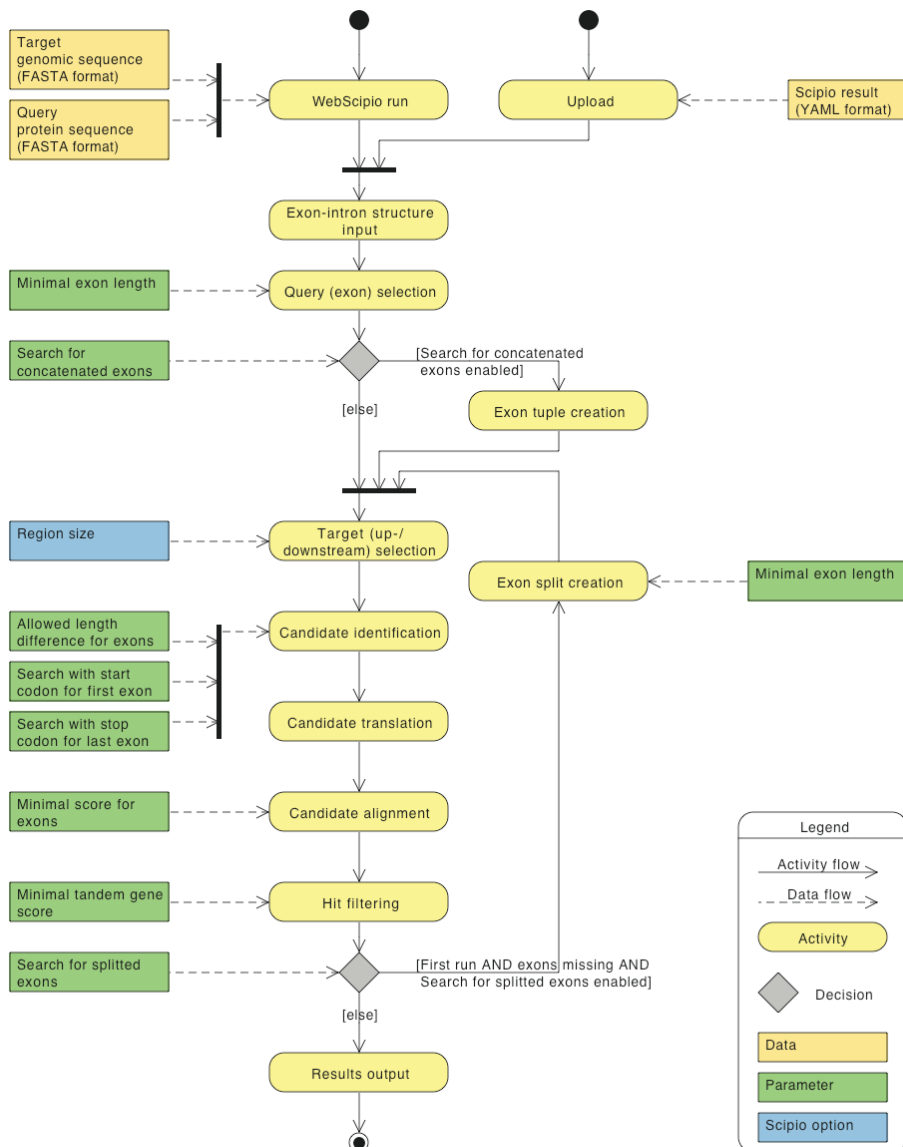


Fig. 1. Activity flow diagram of the search for tandem gene duplications: The activity diagram shows the processing steps of the search algorithm and the influence of the parameters on each step. The run starts with an exon-intron gene structure determined by Scipio. Based on the chosen parameters the exons and up- and downstream regions are selected and searched for candidate exons of gene duplicates. The candidates are processed and filtered. These steps are repeated for exons that have not been found. Those exons are splitted and the search is repeated with fragments. In the end, the algorithm outputs the exon-intron structure of the original gene and all gene duplicates.

### 2.1.1 Query and target selection

The next steps are the selection of the query and the target for the search. All exons, which are longer than a minimal length, are selected as query. The minimal length can be adjusted by the *minimal exon length* parameter, which is given in number of amino acids coded by the exon. In addition, the algorithm is able to generate exon tuples by the fusion of neighbouring exons to one exon. This means that all pairs (2-tuples) of consecutive exons, triplets (3-tuples), 4-tuples, 5-tuples, up to all exons are concatenated and used as query exons. This option can be enabled by the *search for concatenated exons* parameter. The nucleotide sequences of the up- and downstream regions of the gene are used as target sequences. The lengths of these sequences are determined by the Scipio parameter *region size* in number of nucleotides. The up- and downstream sequences are scanned in forward and reverse direction. For the reverse strand the reverse complements of the given target sequences are created.

### 2.1.2 Candidate identification

The query and target selection steps are followed by the search for exon candidates in the target sequences. The search algorithm assumes that exons of gene duplications have a similar length, share sequence similarity, are translated in the same reading frame and have conserved splice sites. Candidate exons are determined in the target sequences for each exon of the original gene and each exon tuple. The target nucleotide sequences are scanned for sequence sections, which do not differ more than a maximal number of nucleotides from the original exon length. This maximal difference is given by the *allowed length difference for exons* parameter in number of amino acids. In addition, the sequence section, which determines an exon candidate, must be flanked by a splice site pattern that corresponds to the introns surrounding the original exon or exon tuple. Allowed splice site patterns for the first two and last two nucleotides of these introns are GT---AG, GC---AG, GG---AG, and AT---AC. The first exon of a gene must start with the start codon ATG and the last exon must be followed by one of the stop codons TAG, TAA, or TGA. To allow searches for partial genes, the algorithm is able to find candidates corresponding to the first and last exon of the gene fragment that share splice site patterns instead of having a start codon or stop codon. This behaviour can be adjusted by the *search with start codon for first exon* and *search with stop codon for last exon* parameters.

### 2.1.3 Candidate translation and alignment

Candidate sequences are translated to amino acids in the same reading frame as the original exon. If a candidate sequence includes a stop codon, the candidate is rejected immediately. The translations of the candidate exons are aligned to the original exon translations by a global alignment algorithm. The *pair\_align* tool of the SeqAn package (Doring et al., 2008) is used for this task. The resulting alignment score is divided by the score resulting from the alignment of the original exon translation to itself. This normalised score makes exons of different lengths and amino acid compositions comparable. Finally, exon candidates having a score lower than the score given by the *minimal score for exons* parameter are rejected.

### 2.1.4 Hit filtering

The resulting candidate hits are filtered. If candidate sequences are overlapping, the lower scoring candidates are rejected. Neighbouring candidate exons are combined to genes if they

are in the same order as the original exons. For each identified tandem gene a score is calculated that reveals how many residues of the original gene were found in the tandem gene duplication. The score is calculated as the number of residues of the original gene that are aligned to residues in the tandem gene duplicate (and not to gaps) divided by the number of all residues of the original gene. The tandem gene duplications that have a low score are rejected. This behaviour can be adjusted by the *minimal tandem gene score* parameter.

### 2.1.5 Exon split run

If exons of a duplicated gene are missing, either in between two neighbouring exons, at the start of the gene or at the end, the search is repeated for these exons by splitting the missing original exons into pieces. The original exon sequences are split in two parts at each nucleotide as long as the smaller part is longer than the minimum exon length. The algorithm scans the intron regions of the duplicated genes that miss exons for candidates corresponding to these exon splits, each composed of two parts. Thus, exons, which are split by an intron in the duplicated gene, are found too. This option can be enabled by the *search for splitted exons* parameter.

### 2.1.6 Results output

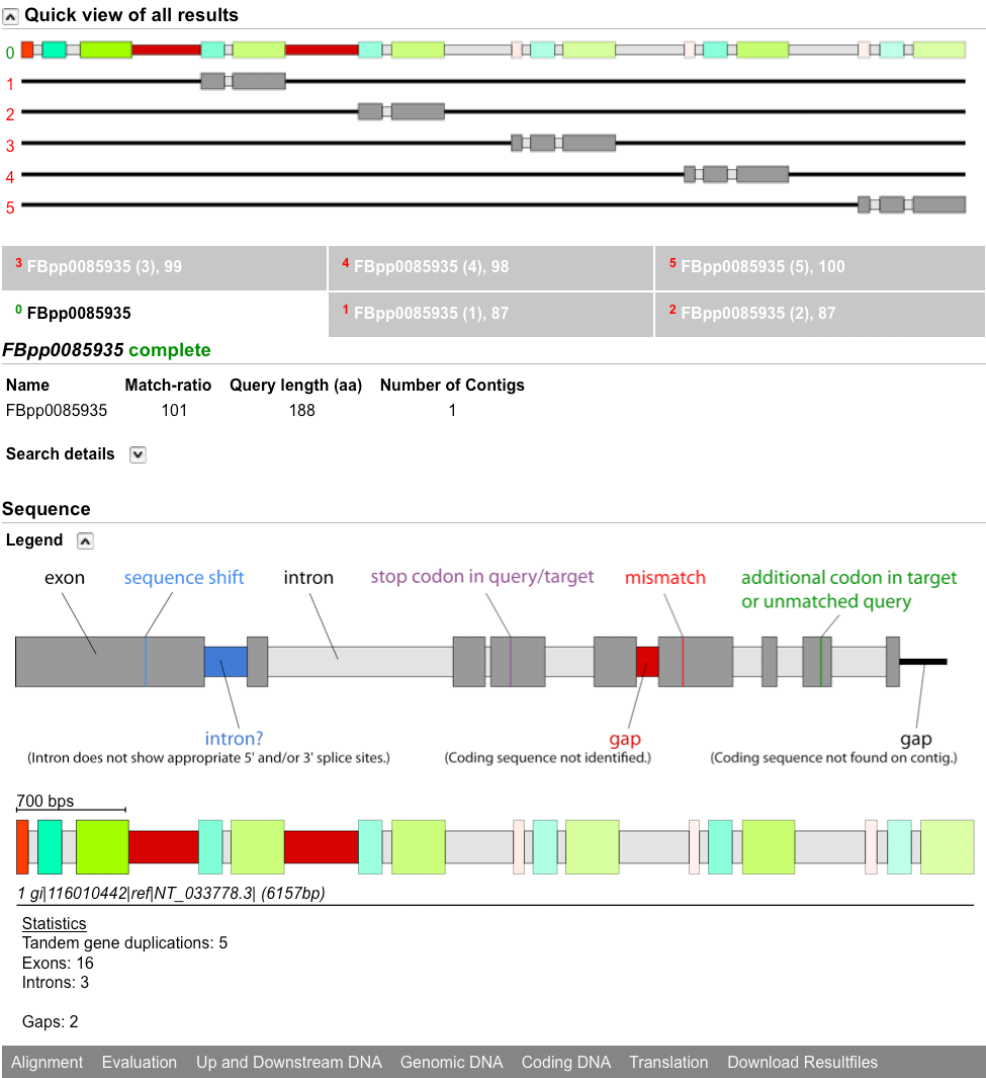
The output of the search algorithm is the exon-intron structure of all identified tandem gene duplications combined in one result, and the exon-intron structure of each duplicated gene alone. For every result a gene structure drawing is shown, as well as several options to further examine gene details like the alignment of the query sequence to the translation of the hit and the hit itself (Fig. 2).

## 2.2 WebScipio integration

The search algorithm is fully integrated into the web interface of WebScipio. The search for tandem gene duplications can be enabled in the Advanced Options section. WebScipio provides an interface to easily set the parameters, suggests default parameters, which will be suitable for most cases, and offers documentation at several help pages and examples. The raw results for the gene cluster can be downloaded all together in one YAML file or the result for each gene of the cluster in a separate file. In addition to the raw data, the SVG figures of the gene structures and FASTA files of the sequences (cDNA, genomic DNA, exons, introns, target translation) are available for download. WebScipio provides an upload option for downloaded YAML files to let the user analyse his results at a later date.

## 3. Results and discussion

WebScipio uses the command-line tool Scipio to reconstruct the gene structures of given protein sequences based on the available eukaryotic genome assemblies. Scipio has been developed for the case that protein sequences and target genome sequence are from the same organism. Nevertheless, Scipio allows several mismatches that might result from sequencing and assembly errors like missing or additional bases, which lead to frame-shifts, or in-frame stop codons that would lead to premature gene stops. Mismatches might also be the result from differences in the source of the protein sequence, which might have been obtained from cDNA libraries of a certain strain, and the specific sequenced strain of the species. To accomplish this task, Scipio relies on BLAT, which is one of the fastest tools available for





query and target gene. If genes are highly conserved in evolution Scipio is able to correctly reconstruct genes in species that diverged hundreds of million years ago. If genes evolve fast Scipio can predict genes only in very related organisms. This behaviour can also be used to predict gene duplicates in the same organism, and is implemented as *multiple results* parameter in the Scipio options. Again, because Scipio relies on BLAT, only those duplicates will be identified that are very similar. An advantage of this option is that Scipio is able to find dispersed as well as tandem duplicates.

In an analysis of the origin of new genes in the *Drosophila* species complex (Zhou et al., 2008) it has been shown that the majority of the constrained functional new genes are dispersed duplicates. In contrast, tandem duplications were found to be young events and to lead to lower survival rates. Thus, tandem duplicates are often pseudogenes most probably because the introduction of frame shifts and in-frame stop codons does not demand too many mutations to destroy the transcription and expression of the new gene. If duplicates are kept in the genome they acquire new functions through neofunctionalization and subfunctionalization by accumulation of many substitutions (Ohno, 1970). Those genes are too divergent to be identified by the *multiple results* option of Scipio. However, although accumulating many substitutions tandem duplicates very often retain the gene structure of the original gene including intron splice sites and reading frames of exons. Occasionally further introns might be introduced or prior existing introns lost because these changes would not destroy transcription and translation. To use this knowledge in tandem gene duplicate identification we developed an algorithm that searches for duplicates of a query sequence based on the restrictions imposed by its gene structure. Every piece of DNA in the up- and downstream region of the original exon that has the same splice sites and shares sequence homology to the original exon, when translated in the same reading frame, is thought to be a candidate for an exon of a duplicated gene. In the case that introns have been lost or gained in the duplicated genes the splice site restrictions apply to the outer borders of the fused or split exons. WebScipio is able to correctly reconstruct the gene structure for a given protein sequence and is thus very suited as starting point for searches for candidate exons of duplicated genes.

To search for tandem gene duplicates an extension to WebScipio was implemented providing several parameters to adjust the search according to users or genome-specific needs. In most cases, however, the standard parameters will provide reasonable and interpretable results. As soon as the search is done, WebScipio shows an overview of the results as small gene structure pictures (Quick View), which reveal the exon regions of the found tandem genes (Fig. 2). For convenient analysis the genomic region comprising the gene structure of the query sequence and the exons of the predicted tandem genes is shown in a combined graph and provided as one YAML file. The exons of the original gene are dark coloured and the corresponding predicted exons have the same but lighter colour. The darkness of the colour relates to the similarity of the predicted exon to the original one. The same colour scheme is used to highlight the various exons in the Alignment view of the genomic regions (Fig. 3). The Alignment view shows the nucleotide sequence of the gene ordered in exons and introns. For every exon the genomic DNA and the corresponding translation are shown, as well as the alignment of the query sequence to the translation.

To demonstrate the application, quality, and limitations of the new algorithm we provide some example searches in the following sections. Tandemly arrayed gene duplicates have several characteristics that need to be considered. Gene duplications can be found on both the forward and the reverse strand. The duplicated genes might contain fused exons or

might contain additional introns. In the case of retroposed genes, which are derived from the reverse transcription and insertion of processed genes, gene duplications do not have

any introns. Although gene duplications are more often found for small genes consisting of one or only a few exons, gene duplicates can also be identified for genes consisting of dozens of exons spanning large genomic regions. Because tandem gene duplicates are defined by being located next to each other in the genome, intergenic regions are expected to be short. This is also the reason why the parameter for bordering the search in up- and downstream regions of the original gene limits this region to 300,000 nucleotides. However, WebScipio cannot exclude that there may be additional genes in-between gene duplicates. An example for such a scenario would be the duplication or multiple duplications of small genomic regions that encode several genes. In most cases we considered examples from the fruit fly *Drosophila melanogaster* and sequences from Flybase (Tweedie et al., 2009), because the corresponding genome is of high quality and the annotation of the genome is already at a very advanced stage. Fragmented genomes, like draft genomes for which only short contigs are available, or chromosome assemblies containing many gaps, are useful to screen for interesting candidates but do not provide the reliability needed for tests of the algorithms quality and limitations. An advanced annotation provides the advantage that genomic locations of most genes have already been identified. Thus the gene order is already established although there might still be errors in the annotation of single exons.

### 3.1 Examples of tandemly arrayed gene duplicates

#### 3.1.1 Gene duplicates on both the forward and the reverse strand

The WebScipio tandem gene duplication extension has been developed to find tandem gene duplications on the forward as well as on the reverse strand in relation to the query gene. The example in Fig. 4 shows five gene duplicates of the *Drosophila melanogaster* heat shock protein 23 gene (Hsp23), which consists of one exon. The first duplicate (Hsp67Bc) and the forth duplicate (Hsp26) in the genomic region are on the reverse strand, the other duplications Hsp22, CG4461, and Hsp27 are in the same reading direction as Hsp23. This search was performed with default parameters except increasing the *allowed length difference for exons* parameter to 30 amino acids. The most divergent gene duplication Hsp67Ba (Table 1), which is encoded in the genomic region between Hsp26 and Hsp23, was not found. This example shows that although the sequence identity is very low between the duplicates and the Hsp23 search sequence (Table 1), five duplicates could be identified. The length difference between Hsp23 and Hsp67Ba was too large so that candidates of the length of Hsp67Ba were not included in the search with the given search parameters.

	Hsp67Bc	Hsp22	CG4461	Hsp26	Hsp67Ba	Hsp23	Hsp27
Length [aa]	199	174	200	208	445	186	213
Identity to Hsp23	0.29	0.31	0.26	0.49	0.15	1.00	0.41
Strand	rev	for	for	rev	rev	for	for

Table 1. Comparison of the length, similarity, and reading direction of the genes of the *Drosophila melanogaster* heat shock protein cluster.

*Drosophila melanogaster* heat shock protein 23 gene and duplicated genes on both strand

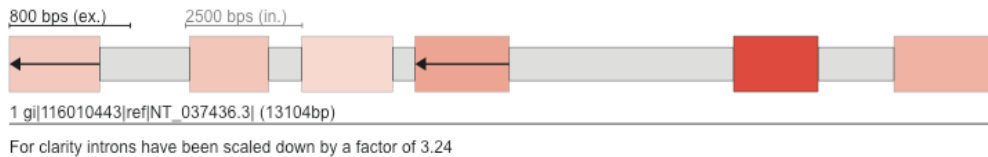


Fig. 4. *Drosophila melanogaster* heat shock protein gene duplicates: The figure shows the duplications found by the algorithm with Hsp23 as query. The genomic region contains, from the left to the right side in the drawing, the identified genes Hsp67Bc, Hsp22, CG4461, Hsp26, the query gene Hsp23, and another gene duplicate Hsp27. Gene duplications on the reverse strand are marked by an arrow in reverse direction.

### 3.1.2 Duplicated exons in six tandemly arrayed genes including a lost intron and a pseudogene

The new algorithm is able to reconstruct tandemly arrayed gene duplications containing many exons and gene duplicates. The *Drosophila melanogaster* CG30047 gene includes 12 exons. Five duplicates of this gene could be identified with the algorithm (Fig. 5, top). In the second duplicated gene an intron loss could be identified. The exons 11 and 12 of CG30047 are translated as one exon in this duplicate (Fig. 5, bottom). To find such lost introns the option to *search for concatenated exons* has been enabled. The third duplicate most probably represents a pseudogene, because exon 11 contains a frame shift and could thus not be found. Other reasons for the frame shift could be sequencing and assembly errors. However, the *Drosophila melanogaster* genome (Adams et al., 2000) is one of the best available and a lot of effort has been spent in the finishing process. Thus, it is more probable that the third duplicate is a pseudogene. Exon 1, which codes for seven amino acids, has low complexity and could therefore only be identified in the second gene duplication by setting the *minimal exon length* parameter to 7 aa.

### 3.1.3 Myosin heavy chain gene duplicates

Mammals encode two clusters of muscle myosin heavy chain genes, one cluster containing the  $\alpha$ - and  $\beta$ -cardiac muscle myosin heavy chain genes (Saez et al., 1987; Weydert et al., 1985), and one cluster containing six skeletal muscle myosin heavy chain genes in the order embryonic, 2a, 2x, 2b, perinatal, and extraocular (Sun et al., 2003; Weydert et al., 1985). These myosin genes consist of 38 exons each. Based on their gene size and number of exons the genes of the muscle myosin gene cluster should be on the upper limit of the complexity of a search for tandem gene duplicates. With the new WebScpio extension all genes of the muscle myosin cluster in *Homo sapiens* could be identified (Fig. 6). For the search the *region size* parameter was set to 300,000 nucleotides and the *minimal score for exons* to 50 %. This example also shows the advantage of the new WebScpio extension compared to the *multiple results* option in Scpio. When searching with the *multiple results* option of Scpio and the 2a gene as starting sequence, mixed genes are found for every additional gene candidate (Fig. 6). Scpio does not know about gene borders and analyses all BLAT hits according to their score. Therefore, Scpio combines the highest scoring hits to gene candidate one (2a), the next highest scoring hits to gene candidate two (2x), and so on. The third gene

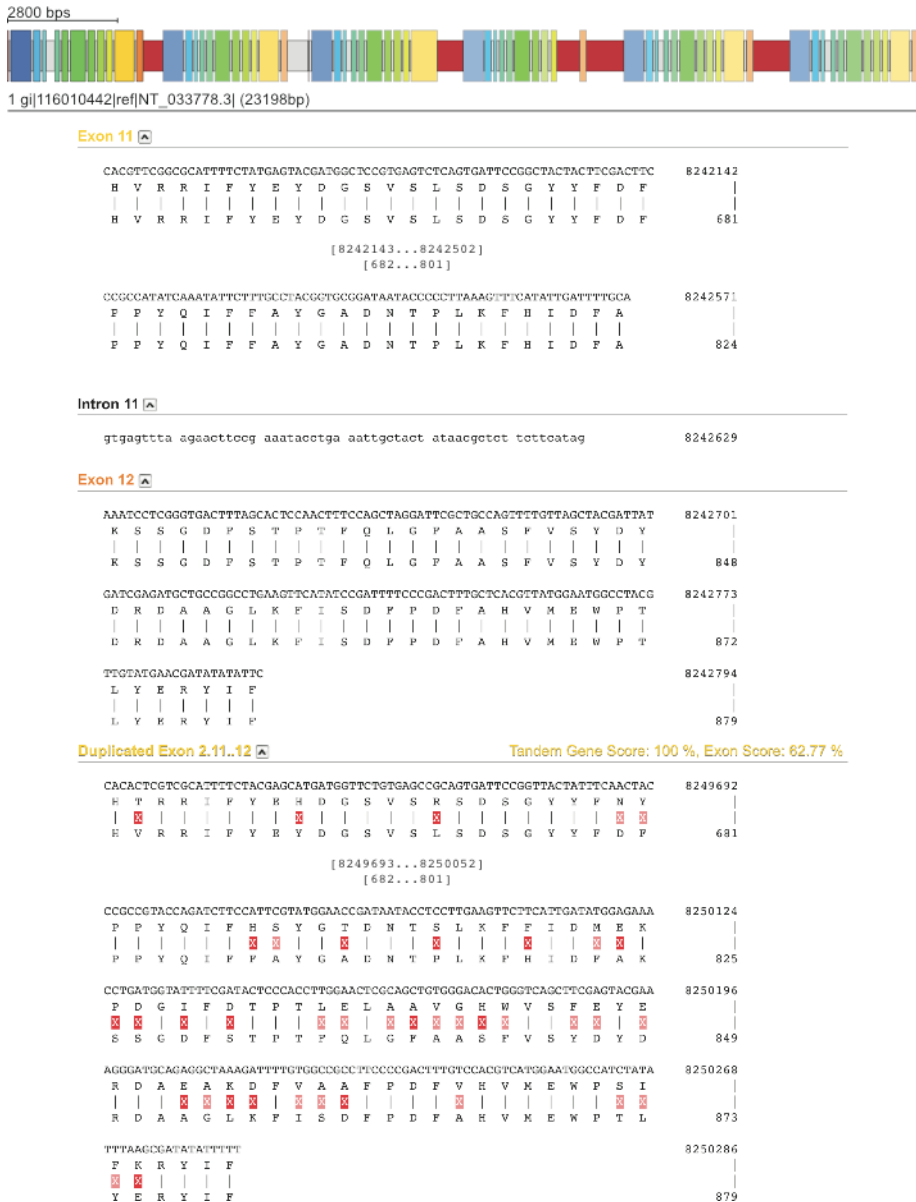
*Drosophila melanogaster* CG30047 gene and duplicates

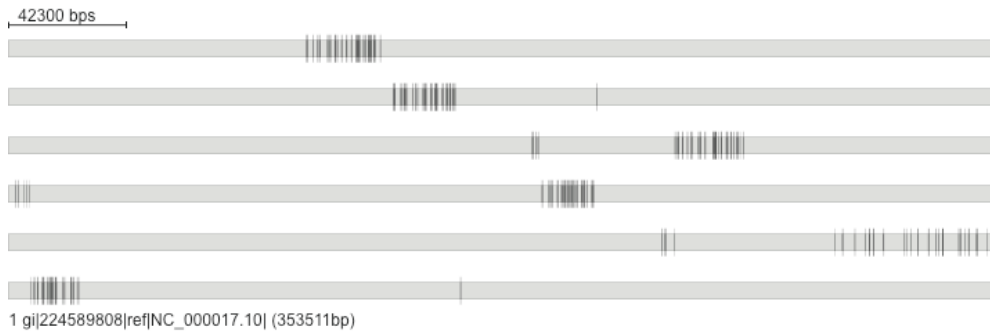
Fig. 5. *Drosophila melanogaster* CG30047: Five gene duplications were found for the CG30047 gene. In the second duplication the intron between exon 11 and 12 was lost as shown in the alignment. The alignment of exon 11 (CG30047) and the alignment of the corresponding region in the duplicated gene were shortened by amino acids 682 to 801 for representation purposes.

### *Homo sapiens* muscle myosin heavy chain gene cluster

Genes found with new algorithm



Genes found with Scipio (*multiple results* parameter enabled)



Genes found with new algorithm (scaled)



For clarity introns have been scaled down by a factor of 9.17

Fig. 6. *Homo sapiens* muscle myosin heavy chain gene cluster: The skeletal muscle myosin heavy chain cluster consists of the genes embryonic, 2a, 2x, 2b, perinatal, and extraocular, from left (5' end) to the right (3' end). The WebScipio search for tandem gene duplicates based on the 2a gene identifies all other genes of the cluster. The Scipio search with the parameter *multiple results* also identifies six gene candidates but only the search sequence (the 2a gene) is found correctly while the other gene candidates consist of fusions of different parts of the other muscle myosin heavy chain genes.

candidate, for example, mainly consists of the exons of the perinatal muscle myosin heavy chain gene, but the N-terminus of the 2b gene has a higher homology to the 2a gene than the N-terminus of the perinatal gene and therefore the 2b N-terminus is combined with the C-terminus of perinatal.

The Nile tilapia *Oreochromis niloticus* contains another type of a muscle myosin heavy chain gene cluster (Fig. 7). Here, two genes (Mhc6 and Mhc13) are encoded on the forward strand, and Mhc7 is encoded on the reverse strand. Nevertheless, WebScipio correctly reconstructed the complete cluster when searching with the Mhc13 gene. When searching with Mhc6 or Mhc7, the small C-terminal exons of the respective other genes could not be identified. These examples demonstrate that WebScipio with the new extension is able to correctly identify arrays of very large and complex genes. For this search the minimal score for exons parameter was set to 30 % and the region size parameter to 50,000 nucleotides.

*Oreochromis niloticus* myosin heavy chains 6, 7 and 13

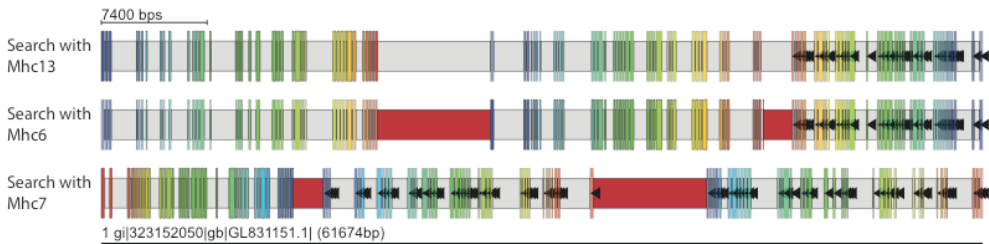


Fig. 7. *Oreochromis niloticus* muscle myosin heavy chain gene cluster: The Nile tilapia contains a cluster of three muscle myosin heavy chain genes (Mhc13, Mhc6, and Mhc7) of which Mhc7 is encoded in the opposite direction. The last exon is too divergent to be identified in most cases. Only when searching with the Mhc13 gene, the tandem genes Mhc6 and Mhc7 are reconstructed completely.

### 3.1.4 Revealing a pseudogene

For the *Drosophila melanogaster* CG3397 gene the first exon is splitted into two exons in the prediction of the gene duplication. For this search the default parameters were used and the option to *search for splitted exons* was enabled. The predicted gene is most probably a pseudogene, because either the predicted intron between the two splitted exons is too short to be spliced, or the exon translation results in a frame shift if both parts are considered as one potential exon. The details are shown in the alignment (Fig. 8).

## 3.2 Examples of non-tandemly arrayed gene duplicates

### 3.2.1 Duplicated gene regions

Tandemly arrayed genes evolve through unequal recombination. In this process not only single genes might be duplicated but small genomic regions containing several genes. The result would be a tandemly arrayed group of genes. Because WebScipio is searching for each gene separately it cannot separate a group of duplicated genes from a tandem array of single genes. An example for duplicated genomic regions is the region in *Drosophila melanogaster* containing genes coding for histones (Fig. 9). The new algorithm identified many duplicates for each of the His1, His2A, His2B, His3, and His4 genes in the *Drosophila* genome. As query the genes CG33825 (His1), CG33826 (His2A), CG33894 (His2B), CG33827 (His3), and CG33893 (His4) were used. The His2B and His4 genes are on the reverse strand in comparison to the other genes. The genes are very similar (some code for the same protein sequence) resulting in alignment scores between 99 % and 100 %. Only two more divergent gene duplicates were found for the His2A gene. The first two gene duplicates of His2A have alignment scores of 79 %.

### 3.2.2 Trans-spliced genes

Tandem gene duplicates and *trans*-spliced genes could evolve through the same gene duplication process during evolution, except that only part of the gene is duplicated instead of the complete gene. The exon-intron structure of tandem gene duplicates and *trans*-spliced genes look very similar, which complicates their differentiation during the process of gene identification. If, for example, the constitutive part of the *trans*-spliced gene consists of only

*Drosophila melanogaster* gene CG3397

Fig. 8. *Drosophila melanogaster* CG3397 gene: A gene duplication could be identified downstream of the CG3397 gene, which, however, most probably is a pseudogene.

one exon while the *trans*-spliced part consists of groups of similar alternative exons the correct reconstruction of the *trans*-spliced gene would not look different compared to a partial reconstruction of a cluster of duplicated genes for which the first (or last) exons were not found because of low similarity. The gene CG1637 of *Drosophila* is a *trans*-spliced gene (McManus et al., 2010). The WebScipio algorithm predicts tandemly arrayed genes for isoform A and B of CG1637, although the first exons of the potential tandem gene candidates were not found (Fig. 10). The close inspection of the three isoforms shows that the predicted exons do not belong to duplicated genes, but to *trans*-spliced variants of the same gene. Another type of problem is demonstrated by the dynein intermediate chain gene of *Drosophila melanogaster*. Here, the dynein intermediate chain gene is annotated as four separate genes (Sdic1, Sdic2, Sdic3 and Sdic4) in Flybase (version of June 24<sup>th</sup>, 2011). The problem is, however, that the real first two exons of the gene are not annotated in Flybase.



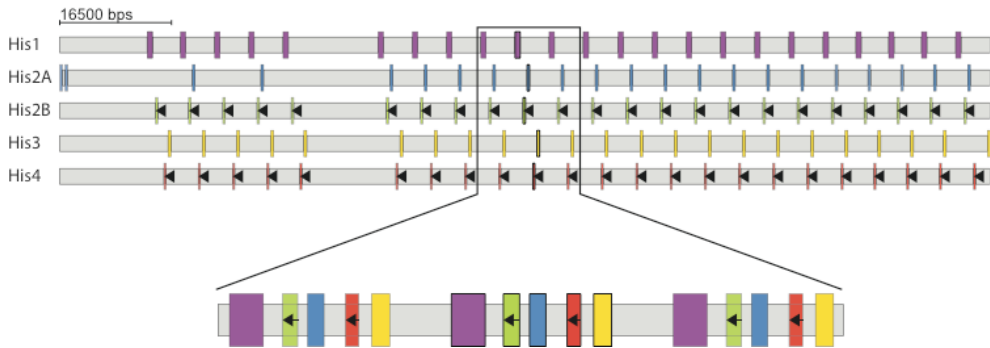


Fig. 9. *Drosophila melanogaster* histones: The results for the separate searches for gene duplicates of the histones His1, His2A, His2B, His3, and His4 are shown. Based on the results of the search for each single gene it is not possible to distinguish between a gene and a genomic region duplication. The results of all searches at the same scale shows that not single genes but a genomic region containing all five histone genes has been duplicated several times.

#### *Drosophila melanogaster* gene CG1637



#### *Drosophila melanogaster* dynein intermediate chain



For clarity introns have been scaled down by a factor of 3.84

Fig. 10. *Drosophila melanogaster* CG1737 gene and *Drosophila melanogaster* dynein intermediate chain: The algorithm identified duplicated exons in the *trans*-spliced CG1737 and dynein intermediate chain genes. The search was done with default parameters and the search for concatenated exons and search for splitted exons options were enabled. To reveal the last and most divergent exon the region size parameter was set to 35,000 nucleotides and the allowed length difference parameter to 30 amino acids for the dynein intermediate chain gene.

The sequence encoded by the true first exons is conserved throughout all major branches of the eukaryotic tree of life that express a cytoplasmic dynein, in chromalveolates, Excavata, and Opisthokonta. In addition, this N-terminal part of the dynein intermediate chain is of high functional importance because it connects dynein to dynactin by interacting with the

dynactin p150 gene. Based on these facts and the found exon order of the genomic region, we expect the gene to be *trans*-spliced (Fig. 10, bottom).

#### 4. Conclusion

Our algorithm provides a method to consistently predict and reconstruct tandemly arrayed gene duplicates. It has been integrated into the web interface of WebScpio allowing the search for gene duplicates of a given query protein sequence in the respective genome assemblies. WebScpio provides access to more than 2300 genome assembly files from more than 650 eukaryotes (July 2011) and is updated as soon as further genome assemblies become available whether from newer versions of already sequenced species or from newly sequenced genomes. The search results are presented in drawings coloured according to the sequence similarity of the gene duplicate to the search sequence, and in several human-readable formats like detailed alignments of the found exons to the genomic DNA. Sequences and figures can be downloaded, as well as the complete raw data for later upload or further computational analysis. The new algorithm is based on the precondition that gene duplicates rather retain the gene structure of the original gene than the sequence. We could show that the new extension to WebScpio is able to correctly predict and reconstruct gene duplicates on both the forward and the reverse strand. Also, the new algorithm is able to correctly reconstruct complicated gene structures spread over hundreds of thousands of nucleotides like the skeletal muscle myosin heavy chain gene cluster in mammals. Gene duplications often accumulate gene function destroying mutations that lead to frame shifts and in-frame stop codons. Those potential pseudogenes are identified by WebScpio but the user has to carefully inspect the results to distinguish between sequencing errors and real pseudogenes. WebScpio cannot distinguish between gene duplicates and duplications of small genomic regions that might encode several genes. Here, WebScpio can identify and reconstruct the duplicates of one gene but does not provide any hints about other genes in the intergenic regions. *Trans*-spliced genes often contain clusters of alternative exons. Those clusters will be identified by WebScpio, but again the user needs to evaluate the results to distinguish between cases of *trans*-spliced genes, where the constitutive part is encoded by just a few exons, or real gene duplications, for which some terminal exons could not be identified because of very low sequence similarity or even assembly gaps. Altogether, WebScpio provides an easy to use way to analyse the genomic region of every gene of interest for the very common event of tandem gene duplication.

#### 5. Acknowledgments

MK has been funded by grant KO 2251/6-1 of the Deutsche Forschungsgemeinschaft. We thank Björn Hammesfahr for fruitful discussions.

#### 6. References

- Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F. et al. (2000). The genome sequence of *Drosophila melanogaster*, *Science*, Vol.287, No.5461, pp. 2185-2195
- Aloni, R., Olender, T. & Lancet, D. (2006). Ancient genomic architecture for mammalian olfactory receptor clusters, *Genome Biol*, Vol.7, No.10, pp. R88

- Babushok, D. V., Ostertag, E. M. & Kazazian, H. H., Jr. (2007). Current topics in genome evolution: molecular mechanisms of new gene formation, *Cell Mol Life Sci*, Vol.64, No.5, pp. 542-554
- Bertrand, D., Lajoie, M. & El-Mabrouk, N. (2008). Inferring ancestral gene orders for a family of tandemly arrayed genes, *J Comput Biol*, Vol.15, No.8, pp. 1063-1077
- Doring, A., Weese, D., Rausch, T. & Reinert, K. (2008). SeqAn an efficient, generic C++ library for sequence analysis, *BMC Bioinformatics*, Vol.9, pp. 11
- Elemento, O., Gascuel, O. & Lefranc, M. P. (2002). Reconstructing the duplication history of tandemly repeated genes, *Mol Biol Evol*, Vol.19, No.3, pp. 278-288
- Garcia-Fernandez, J. (2005). The genesis and evolution of homeobox gene clusters, *Nat Rev Genet*, Vol.6, No.12, pp. 881-892
- Goto, N., Prins, P., Nakao, M., Bonnal, R., Aerts, J. & Katayama, T. (2010). BioRuby: Bioinformatics software for the Ruby programming language, *Bioinformatics*
- Hahn, M. W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates, *J Hered*, Vol.100, No.5, pp. 605-617
- Keller, O., Odronitz, F., Stanke, M., Kollmar, M. & Waack, S. (2008). Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species, *BMC Bioinformatics*, Vol.9, pp. 278
- Kent, W. J. (2002). BLAT--the BLAST-like alignment tool, *Genome Res*, Vol.12, No.4, pp. 656-664
- Li, W. H., Yang, J. & Gu, X. (2005). Expression divergence between duplicate genes, *Trends Genet*, Vol.21, No.11, pp. 602-607
- Long, M., Betran, E., Thornton, K. & Wang, W. (2003). The origin of new genes: glimpses from the young and old, *Nat Rev Genet*, Vol.4, No.11, pp. 865-875
- Massingham, T., Davies, L. J. & Lio, P. (2001). Analysing gene function after duplication, *Bioessays*, Vol.23, No.10, pp. 873-876
- McManus, C. J., Duff, M. O., Eipper-Mains, J. & Graveley, B. R. (2010). Global analysis of trans-splicing in *Drosophila*, *Proc Natl Acad Sci U S A*, Vol.107, No.29, pp. 12975-12979
- Nei, M. & Roychoudhury, A. K. (1973). Probability of fixation and mean fixation time of an overdominant mutation, *Genetics*, Vol.74, No.2, pp. 371-380
- Odronitz, F., Pillmann, H., Keller, O., Waack, S. & Kollmar, M. (2008). WebScipio: an online tool for the determination of gene structures using protein sequences, *BMC Genomics*, Vol.9, pp. 422
- The Official YAML Web Site, (2011). Available from <http://www.yaml.org/>
- Ohno, S. (1970). Evolution by Gene Duplication, Berlin, *Springer*
- Prototype JavaScript framework: Easy Ajax and DOM manipulation for dynamic web applications, (2011). Available from <http://www.prototypejs.org>
- purzelrakete's working at master - GitHub, (2011). Available from <http://github.com/purzelrakete/working>
- Quijano, C., Tomancak, P., Lopez-Marti, J., Suyama, M., Bork, P., Milan, M., Torrents, D. & Manzanares, M. (2008). Selective maintenance of *Drosophila* tandemly arranged duplicated genes during evolution, *Genome Biol*, Vol.9, No.12, pp. R176
- Ruby on Rails, (2011). Available from <http://rubyonrails.org>
- Ruby Programming Language, (2011). Available from <http://www.ruby-lang.org/>

- Saez, L. J., Gianola, K. M., McNally, E. M., Feghali, R., Eddy, R., Shows, T. B. & Leinwand, L. A. (1987). Human cardiac myosin heavy chain genes and their linkage in the genome, *Nucleic Acids Res*, Vol.15, No.13, pp. 5443-5459
- script.aculo.us - web 2.0 javascript, (2011). Available from <http://script.aculo.us>
- Shoja, V. & Zhang, L. (2006). A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat, *Mol Biol Evol*, Vol.23, No.11, pp. 2134-2141
- Sun, Y. M., Da Costa, N. & Chang, K. C. (2003). Cluster characterisation and temporal expression of porcine sarcomeric myosin heavy chain genes, *J Muscle Res Cell Motil*, Vol.24, No.8, pp. 561-570
- Tokyo Cabinet: a modern implementation of DBM, (2011). Available from <http://fallabs.com/tokyocabinet/>
- tra's spawn at master - GitHub, (2011). Available from <http://github.com/tra/spawn>
- Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R. et al. (2009). FlyBase: enhancing Drosophila Gene Ontology annotations, *Nucleic Acids Res*, Vol.37, Database issue, pp. D555-559
- W3C SVG Working Group, (2011). Available from <http://www.w3.org/Graphics/SVG/>
- Weydert, A., Daubas, P., Lazaridis, I., Barton, P., Garner, I., Leader, D. P., Bonhomme, F., Catalan, J., Simon, D., Guenet, J. L. et al. (1985). Genes for skeletal muscle myosin heavy chains are clustered and are not located on the same mouse chromosome as a cardiac myosin heavy chain gene, *Proc Natl Acad Sci U S A*, Vol.82, No.21, pp. 7183-7187
- Zhang, J. (2003). Evolution by gene duplication: an update, *Trends Ecol Evol*, Vol.18, pp. 292-298
- Zhang, J. & Nei, M. (1996). Evolution of Antennapedia-class homeobox genes, *Genetics*, Vol.142, No.1, pp. 295-303
- Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S. & Wang, W. (2008). On the origin of new genes in Drosophila, *Genome Res*, Vol.18, No.9, pp. 1446-1455

# The LRR and TM Containing Multi-Domain Proteins in Arabidopsis

Felix Friedberg

*Howard University Medical School Washington, DC*

*USA*

## 1. Introduction

Thousands of different multi-domain proteins, each the product of one separate specific gene, which exhibit one or multiple transmembrane (TM) domains, are expressed in Arabidopsis as well as in Homo sapiens (Sallman-Almen et al.,2009). These molecules may carry one or multiple TM domains very close to the amino or carboxyl terminus or even dispersed throughout the molecule. Thus one may conclude that the TM domain existed prior to the branching of plants and metazoa. In both organisms, a very small percent of the TM domain containing proteins additionally possess multiple leucine rich repeats (LRR) domains. These domains seem to transmit ligand perception. So one should presume that such domains also existed prior to the branching of these two forms of life. (A caveat, however, is in order: Many of the combinations of the various domains which one finds in plants are not present in the same combination in animals and vice versa.). As demonstrated in this paper, subsequently, during evolution, on several occasions, the genes for these multi-domain proteins duplicated and during this process they were often altered slightly to allow generation of proteins that could provide new specific functioning (i.e. they underwent neofunctionalization).

Unlike the “adaptive immune” system which exists in animals but not in plants, an “innate immune” system is present in all multicellular organisms (animals and plants). This latter system operates by way of receptors: the Toll-like receptors (TLRs) (first identified in Drosophila) which bind lipopolysaccharides (endotoxins). In Drosophila, these receptors not only activate innate immunity; they also act in dorsal-ventral specifications. When one compares these molecules as they are encoded e.g. in Arabidopsis (Dangl & Jones, 2001) with those present in Homo Sapiens, one finds that in animals – but not in plants – these receptor proteins possess a cysteine rich domain just prior to entering the membrane and also a signaling domain (which is not a protein kinase but which acts as a docking site) on the backside of the TM domain. TLRs target “pathogen associated molecular patterns” (PAMPs) by way of the LRRs domains.

There are, however, besides of TLRs, many other multi-domain proteins that contain both TM and LRR domains, encoded in Arabidopsis and also in Homo sapiens. Many exist in both of these two species but some of them are present only in one or in the other. Below we list more than a hundred of these various proteins (each expressed from an individual gene) present in Arabidopsis. The TM domain is an about 22 AA residue domain and the LRR is an about 20-29 residue domain (which contains about 6 Leu). Both domains are present in proteins that

participate in protein-protein interactions but which have different functions and cellular locations. In each case, the protein is presented below, in an order guided by a Clustal X arrangement (<http://www.clustal.org/>) and is labeled by its NCBI (<http://www.ncbi.nlm.nih.gov/protein>) protein identification number, followed by its chromosomal locus tag, in diagram form as given by SMART (<http://smart.embl-heidelberg.de>).

### 1.1 Proteins with extracellular multiple LRR domains, a single TM domain (blue rectangle) positioned close to the carboxyl terminus and a short cytoplasmic tail



NP\_199740;AT5G49290; Protein Binding



NP\_177558;AT1G74180; Receptor Like Protein 14; Protein Binding



NP\_180117;AT2G25470; Receptor Like Protein 21; Protein Binding  
(NP\_199740, NP\_177558 and NP\_180117 display 60% AA identity.)



NP\_177559;AT1G74190; Receptor Like Protein 15; Protein Binding



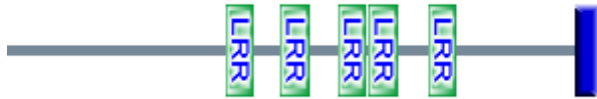
NP\_177557;AT1G74170; Receptor Like Protein 13; Protein Binding



NP\_190892;AT3G53240; Receptor Like Protein 45; Protein Binding



NP\_176115;AT1G58190; Receptor Like Protein 9, Protein Binding



NP\_178125;AT1G80080; TMM (Too Many Mouths); Protein Binding/Receptor. (Note: This protein promotes cell fate progression in stomatal development of stems (Bhave et al.,2009)).



P\_176717;AT1G65380; Clavata 2; Protein Binding /Receptor Signaling Protein. (This protein forms a distinct CLE binding receptor complex regulating stem cell specification (Guo et al.,2010)).



NP\_188941; AT3G23010; Receptor Like Protein 36 (Disease Resistance Protein)



NP\_187188; AT3G05370; Receptor Like protein 31; Protein Binding



NP\_177296; AT1G71400; Receptor Like Protein 12; Protein Binding



NP\_567412;AT4G13920; Receptor Like Protein 50; Protein Binding



NP\_197963; AT5G25910; Receptor Like Protein 52; Protein Binding (Note: It is a putative disease resistance protein induced e.g. by chitin oligomers (Ramonell et al., 2005)).



NP\_187719; AT3G11080; Receptor Like Protein 35; Protein Binding



NP\_187712; AT3G11010; Receptor Like Protein 34; Protein Binding



NP\_189531; AT3G28890; Receptor Like Protein 43; Protein Binding



NP\_187217; AT3G05660; Receptor Like Protein 33; Protein Binding



NP\_179112; AT2G15080; Receptor Like Protein 19; Protein Binding

Note: It is possible that the primary function of most – if not of all – of the proteins listed above is to provide disease resistance.



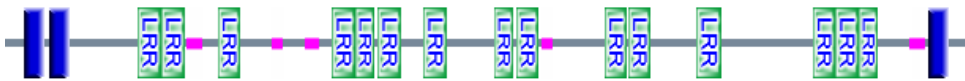
## 1.2 Proteins with one single TM domain positioned at the amino terminus and another one close to the carboxyl terminus



NP\_198058:AT5G27060; Receptor Like Protein 53; protein binding



NP\_187216;AT3G05650; Receptor Like Protein 32; protein binding



NP\_175225;AT1G47890; Receptor Like Protein 7; protein binding. Note: NP\_175225 possesses two TM domains at the AA terminus. ( NP\_198058, NP\_187216 and NP\_175225 exhibit 44 % AA identity.)



NP\_188953;AT3G23120; Receptor Like Protein 38; protein binding:



NP\_188952;AT3G23110; Receptor Like Protein 37; protein binding



NP\_187187;AT3G05360; Receptor Like Protein 30; protein binding



NP\_177295;AT1G71390; Receptor Like Protein 11; protein binding



NP\_180864;AT2G33050; Receptor Like Protein 26; protein binding . Notice loss of carboxyl terminal TM.



NP\_173167;AT1G17240; Receptor Like Protein 2; protein binding



NP\_173168;AT1G17250; Receptor Like Protein 3; protein binding

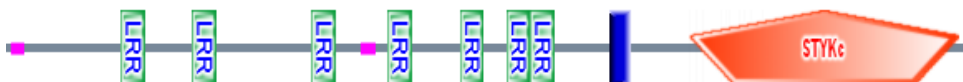
## 2. Receptor Like Kinases (RLKs)

### 2.1 Proteins with extracellular N-terminal LRR domains, a TM domain and an intracellular STYKc (tyrosine kinase catalytic) domain which is close to the C-terminus

These proteins are expected to play roles in intercellular communication during tissue identity maintenance and regulation of development. Still some of them have been shown to play a role in innate immunity. Interaction with the extracellular LRR domains results in activation of the intracellular STYK catalytic domain. Then, the activated receptor, in response, catalyses the phosphorylation of tyrosine residues in intracellular proteins.



NP\_191196;AT3G56370; Leucine rich repeat transmembrane protein kinase. (Inflorescence and root apices receptor-like kinase (IRK). (Hattan et al., 2004)



NP\_195809;AT5G01890; Leucine rich repeat transmembrane protein kinase



NP\_189443;AT3G28040; Leucine rich repeat transmembrane protein kinase



NP\_172708;AT1G12460; Leucine rich repeat transmembrane protein kinase



NP\_193747;AT4G20140; Leucine rich repeat transmembrane protein kinase.

(GASSHO 1).



NP\_199283;AT5G44700; Leucine rich repeat transmembrane protein kinase.

(GASSHO 2). (GASSHO 1 and 2 show 76% AA homology. They are required for the formation of normal epidermal surface during embryogenesis (Tsuwamoto et al., 2008).



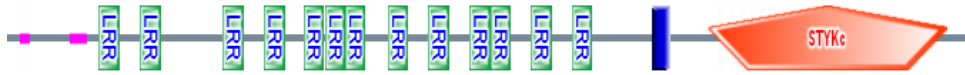
NP\_201371;AT5G65700; Leucine rich repeat transmembrane protein kinase. (BAM1/ Barely any meristem 1). Bam receptors regulate stem cell specification and organ development through interactions with Clavata signaling (DeYoung & Clark, 2008).



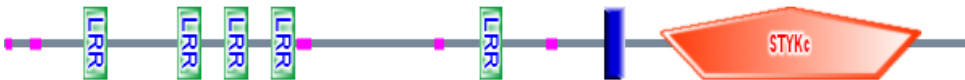
NP\_190536;AT3G49670; Leucine rich repeat transmembrane protein kinase. (Bam2)



NP\_193760;AT4G20270; Leucine rich repeat transmembrane protein kinase (Bam3). BAM 1,2 and 3 exhibit 49% AA identity). Notice the presence of a second TM at the amino terminus of Bam 3 which evolution diluted for Bam 1 and 2.



NP\_180875;AT2G33170; Leucine rich repeat transmembrane protein kinase



NP\_174166;AT1G28440; ATP binding/protein serine/threonine kinase (HAESA-Like1)



NP\_194578;AT4G28490; ATP binding/protein serine/threonine kinase (HAESA)



NP\_201372;AT5G65710; ATP binding / protein serine/threonine kinase (HAESA-Like2) The three HAESA proteins function in developmentally regulated floral organ abscission (Jinn et al., 2000). (NP\_174166, NP\_194578 and NP\_201372 exhibit 37% AA identity.)



NP\_197965;AT5G25930; Leucine rich repeat family protein kinase



NP-199777;AT5G49660; Leucine rich repeat transmembrane protein kinase



NP\_188604;AT3G19700; ATP binding / protein kinase (Haiku 2) Haiku 2 may play a role in the determination of seed size by endosperm (Garcia et al., 2003).



NP\_178330;AT2G02220; ATP binding / protein kinase (Phytosulfokine (PSK) Receptor 1).PSK may play a role in pathogen or herbivore interactions (Loivamaki et al.,(2010) (NP\_188604 and NP\_178330 exhibit 29.3% AA identity.)



NP\_196311;AT5G06940; Leucine rich repeat family protein



NP\_190342; AT3G47580; Leucine rich repeat transmembrane protein kinase



NP\_190293; AT3G47090; Leucine rich repeat transmembrane protein kinase



NP\_566892; AT3G47570; Leucine rich repeat transmembrane protein kinase



NP\_190295; AT3G47110; Leucine rich repeat transmembrane protein kinase



NP\_197548; AT5G20480; ATP binding / protein kinase (EF-TU receptor) It recognizes bacterial PAMP EF - Tu. (Zipfel, et al., 2006).



NP\_199445;AT5G46330; ATP binding / protein kinase (Flagellin sensitive 2 (FLS) ).It recognizes bacterial flagellin and evokes plant innate immunity (Lu, et al., 2010)



NP\_180150;AT2G25790: Leucine rich repeat / transmembrane protein kinase



NP\_178304;AT2G01950; BRI1-like2 ATP binding / protein serine-threonine kinase. This kinase interacts with vascular-specific adaptor proteins to influence leaf venation. (Ceserani et al., 2009.) Note absence of amino terminal TM.



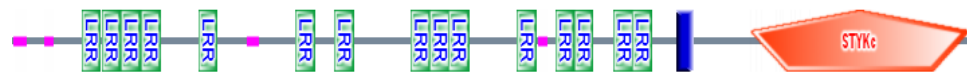
NP\_187946;AT3G13380; BRI1-like3 ATP binding / protein serine-threonine kinase. Note absence of amino terminal TM.



NP\_195650;AT4G39400; Brassinosteroid insensitive 1 (BRI1) ATP binding / protein serine-threonine kinase. (Brassinosteroids regulate plant development by way of a signal transduction pathway involving BRI1 and BAK1 transmembrane receptor kinases (Wang et al., 2008).



NP\_175957;AT1G55610; BRI1-like1 ATP binding / protein serine threonine kinase.



NP\_196345;AT5G07280; EMS1 (Excess Microsporocytes 1) Transmembrane receptor protein kinase. (This protein controls somatic and reproductive cell fates in the anther of Arabidopsis.) (Zhao et al., 2002) Note absence of the amino terminal TM.



NP\_178330;AT2G02220; Phytosulfokine receptor (PSK R1) ATP binding / protein serine-threonine kinase. (PSK represents a class of hormones that affects cellular longevity and growth ( Matsubayashi et al., 2006).



NP\_200200;AT5G53890; Leucine-rich repeat transmembrane protein kinase



NP\_181713;AT2G41820; Leucine-rich repeat transmembrane protein kinase



NP\_174702;AT1G34420; Leucine rich repeat transmembrane protein kinase



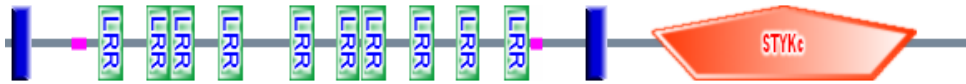
NP\_177694;AT1G75640; Leucine rich repeat transmembrane protein kinase



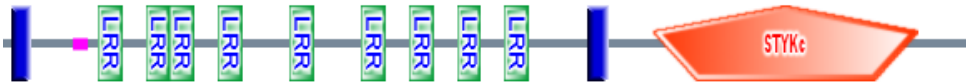
NP\_193826;AT4G20940, Leucine rich repeat transmembrane protein kinase

Note: For molecules such as NP\_172468 SMART does not recognize any LRR regions. Hence, these molecules are not listed here.

**2.2 The members below possess a TM domain at the amino terminus and a second one in front of the STYKc domain:**



NP\_201029;AT5G62230; Erecta - like 1 (ERL1); kinase. Mediates morphological alterations. (Uchida et al.;2011).



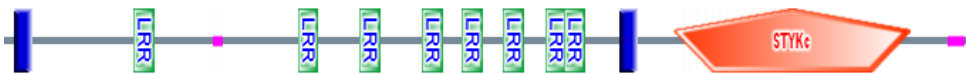
NP\_196335;AT5G07180; Erecta - like 2 (ERL 2); kinase



NP\_173217;AT1G17750;Leucine rich repeat / transmembrane protein kinase. PEP receptor 2 (PEPR2). The amino terminal TM is lost (overly mutated).



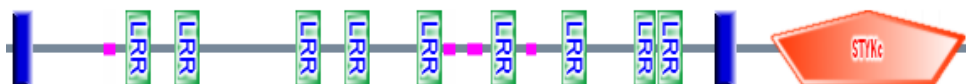
NP\_177451; AT1G73080; ATP binding / protein serine/ threonine kinase, PEP1 receptor (PEPR1). (PEP1 and 2 are implicated in Arabidopsis development and immunity) (Postel, et al.;2010). (PEPR1 and PEPR2, both perceive the existence of an endogenous danger signal Peptide 1 when such a peptide is present.) (Krol et al.;2010). (NP\_173217 and NP\_177451 exhibit 70% AA identity.)



NP\_199705;AT5G48940 Leucine rich repeat transmembrane protein kinase



NP\_189066;AT3G24240; Leucine rich repeat transmembrane protein kinase



NP\_200415;AT5G56040; Leucine rich repeat transmembrane protein kinase





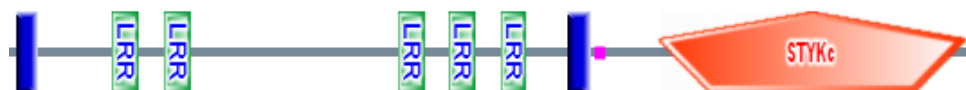
NP\_201198; AT5G63930; Leucine rich repeat transmembrane protein kinase



NP\_174809; AT1G35710; Leucine rich repeat transmembrane protein kinase



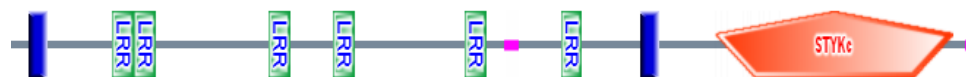
NP\_192625; AT4G08850; Leucine rich repeat transmembrane protein kinase

NP\_200956; AT5G61480; Leucine rich repeat transmembrane protein kinase..PXY receptor.  
(This receptor like-kinase is essential for polarity during plant vascular tissue development.  
(Fisher & Turner, 2007))

NP\_176483; AT1G62950; Leucine rich repeat transmembrane protein kinase



NP\_565084; AT1G74360; Leucine rich repeat transmembrane protein kinase

NP\_177374; AT1G72300; Leucine rich repeat transmembrane protein kinase, (This protein  
"perceives" phytosulfokines (Amano et al., 2007)).

NP\_190742; AT3G51740; Inflorescence Meristem Receptor-Like Kinase (IMK2)



NP\_201529;AT5G67280; Receptor-Like Kinase (RLK)

The Strubbelig Receptor Family Kinases:



NP\_196300;AT5G06820; (SRF2) Strubbelig receptor family 2; kinase. An amino terminal TM is lost. SRFs affect the formation and shape of several organs by influencing cell morphogenesis, the orientation of the division plane and cell proliferation. (Chevalier et al., 2005).



NP\_566444;AT3G13065; (SRF4) Strubbelig receptor family 4; kinase



NP\_178019;AT1G78980; (SRF5) Strubbelig receptor family 5, kinase . The amino terminal TM is lost.



NP\_175777;AT1G53730; (SRF6) Strubbelig receptor family 6; kinase



NP\_188052;AT3G14350; (SRF7) Strubbelig receptor family 7; kinase.. The amino terminal TM is lost.

Not shown:

NP\_565489;AT2G20850; (SRF1) Strubbelig receptor family 1; kinase. All LRR domains are lost. (SMART).

NP\_192248;AT4G03390; (SRF3) Strubbelig receptor family 3; kinase. The amino terminal TM is duplicated. All LRR domains are lost. (SMART).

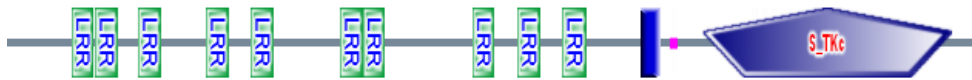
NP\_193944;AT4G22130; (SRF8) Strubbelig receptor family 8, kinase. All LRR domains are lost.(SMART).

Note: For molecules such as NP\_177363 SMART does not recognize any LRR regions. Hence, these molecules are not listed here.

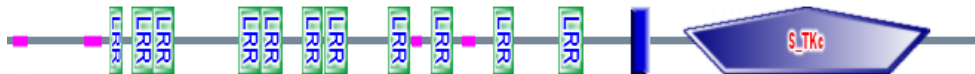
### 3. Proteins that contain an S\_TKc domain: (AA homology is 44% for S\_TKc vs. STYKc.)



NP\_173166;AT1G17230; Leucine rich repeat / transmembrane protein kinase



NP\_174673;AT1G34110; Leucine rich repeat / transmembrane protein kinase



NP\_567748;AT4G26540; Leucine rich repeat / transmembrane protein kinase



NP\_195341;AT4G36180; Leucine rich repeat / transmembrane protein kinase

#### 4. Conclusion

Gene duplication is rampant. It should be noted that the products depicted above show considerable variation both in number of, and distances between, the LRR domains. Where, when and how do the genes listed here function? For most of them there is no evidence that they deliver innate immunity in the plant. Of course, some of them do! For many the function is not known. Gene duplications, carrying amino acid changes resulting from mutations, often end in neofunctionalization even though duplicate genes may also merely provide tissue specific expression for the original ancestral gene. Subsequent alternate splicing of genes, in turn, might also give new roles to the genes. But if domains are the units that built proteins, then domain shuffling provides a more efficient source for expressed gene versatility: (Thereby, nature promotes evolution of disparate proteins for novel functions.) Most of the genes listed in this chapter certainly exist because of duplication. These genes could be grouped further, however, because another domain had first been added for the projected protein molecules, long before the gene duplications occurred. Possibly, domains represent the evolutionary building blocks for all proteins. At present we can only speculate as to the mechanism of such random multi-domain protein formation. Were transposons involved? (Retroproression, the process that is responsible for pseudogene formation, possibly could have also facilitated the creation of new disparate proteins!). Specific domain combinations might have been built randomly – maybe sometimes just once during the evolution of an organism – and then sometimes only to be rearranged during duplication or even to loose domains by mutating them away thereafter. (.See e.g..the Strubbelig family members 1-8.)

#### 5. References

- Amano Y, Tsubouchi H, Shinohara H. et al. Tyrosine-sulfated glycopeptides involved in cellular proliferation and expansion in Arabidopsis, PNAS (2007) 104: 18333-8.
- Bhave NS, Velez KM, Nadeau JA, et al. Mutations of TMM causes stomatal patterning defects in leaves and eliminates stomata formation in stems, Planta (2009) 229: 357-67.
- Ceserani T, Trofka A, Gandora N, et al. VH1/BRL2 receptor-like kinase interacts with vascular-specific adaptor proteins VIT and VIK to influence leaf venation, Plant J (2009) 57: 1000-14.
- Chevalier D, Batoux M, Fulton L, et al. Strubbelig defines a receptor kinase mediating signaling pathway regulating organ development in Arabidopsis, PNAS (2005) 102: 9074-79.
- Clark SE, Williams RW, Myerowitz EM, The Clavata 1 gene encodes a putative receptor kinase that controls shoot and floral meristem size in Arabidopsis, Cell (1997) 89:575-85.
- Dangl JL, Jones JDG, Plant pathogens and integrated defence responses to infection, Nature (2001) 411: 826-33.
- DeYoung BJ, Clark SE, Signalling through the Clavata 1 Receptor complex, Plant Mol Biol (2001) 46:505-13.

- Fisher K, Turner S, PXY a receptor -like kinase essential for maintaining polarity during plant vascular-tissue development, *Curr Biol* (2007) 17: 1061-6
- Garcia D, Saingery V, Chambrier P, et al., Arabidopsis haiku mutants reveal new controls of seed size by endosperm, *Plant Physiol.* (2003) 131: 1661-70.
- Gomez-Gomez L, and Boller T, FLS2: A LRR receptor like kinase involved in the perception of the bacterial elicitor Flagellin in Arabidopsis, *Mol Cell* (2000) 5, 1003-11.
- Guo Y, Han L, Hymes L, et al., Clavata 2 forms a distinct CLE binding receptor complex regulating Arabidopsis stem cell specification, *The Plant Journal* (2010) 63: 889-900.
- Hattan J, Kanamoto H, Takemura M, et al., Molecular characterization of the cytoplasmic interacting protein of the receptor kinase IRK expressed in the inflorescence and root apices of Arabidopsis, *Biosci. Biotechnol. Biochem.* (2004) 68:2598-606.
- Jinn T-L, Stone JM, and Walker JC, HAESA, An Arabidopsis leucine-rich repeat receptor kinase controls floral organ abscission, *Genes, Dev* (2000) 14:108-17.
- Krol E, Mentzel T, Chincilla D, et al., Perception of the Arabidopsis danger signal peptide 1 involves the pattern recognition receptor AtPEPR1 and its close homologue AtPEPR2, *J Biol Chem* (2010) 285:13471-9.
- Loivamaki M, Stuhrwohltdt N, Deeken R, et al. A role for PSK signaling in wounding and microbial interactions in Arabidopsis. *Physiol Plant* (2010) 139:348-57.
- Lu D, Wu S, Gao X, Zhang Y, et al., A receptor-like cytoplasmic kinase, BIK1, associates with a flagellin receptor complex to initiate plant innate immunity, *PNAS* (2010) 107:496-501.
- Matsubayashi Y, Shinohara H, and Ogawa M, Identification and functional characterization of phytosulfokine receptor using a ligand-based approach, *Chem Rec* (2006) 6:356-64.
- Postel S, Kufner I, Beuter C, et al.. The multifunctional leucine-rich repeat receptor kinase BAK1 is implicated in Arabidopsis development and immunity, *European J Cell Biol* (2010) 89: 169-.74.
- Ramonell K, Berrocal-Lobo M, Koh S, et al. Loss-of-function mutations in chitin responsive genes show increased susceptibility to the powdery mildew pathogen *Erysiphe chioracearum*. *Plant Physiol.* (2005) 138: 1027-36.
- Sallman-Almen M, Nordstrom KJ, Fredriksson R, et al.. Mapping the human membrane proteome. *BMC Biology* 2009
- Tori KU, Mitsukawa N, Oosumi T, et al. The Arabidopsis Erecta gene encodes a putative receptor kinase with extracellular leucine rich repeats, *Plant cell* (1996) 8: 735-46.
- Tsuwamoto R, Fukuoka H, Takahata Y, GASSHO1 and GASSHO2 encoding a putative LRR repeat TM-type receptor kinase are essential for the normal development of the epidermal surface of Arabidopsis embryos, *The Plant J.* (2008) 54: 32-42.
- Uchida N, Igan K, Bogenschutz NL, et al. Arabidopsis ERECTA-family receptor kinases mediate morphological alterations stimulated by activation of NB-LRR type UNI proteins, *Plant Cell Physiol.* (2011) in Press

- Zhao DZ, Wang GF, Speal B, Ma H, The excess microsporocytes 1 gene encodes a putative leucine rich repeat receptor protein kinase that controls somatic and reproductive cell fates in the Arabidopsis anther, *Genes Dev.* (2002) 16:2021-31.
- Zipfel C, Kunze G, Cinchilla D, et al. Perception of the bacterial PAMP EF-TU by the receptor EFR restricts Agrobacterium mediated transformation, *Cell* (2006) 125:749-60.

# Partial Gene Duplication and the Formation of Novel Genes

Macarena Toll-Riera<sup>1</sup>, Steve Laurie<sup>1</sup>, Núria Radó-Trilla<sup>1</sup> and M.Mar Albà<sup>1,2</sup>

<sup>1</sup>*Evolutionary Genomics Group, Biomedical Informatics Programme, Universitat Pompeu Fabra (UPF) - Institut Municipal d'Investigació Mèdica (IMIM)*

<sup>2</sup>*Catalan Institution for Research and Advanced Studies (ICREA), Barcelona Spain*

## 1. Introduction

The publication of the first fully sequenced genomes represented a landmark in the biological sciences. The comparison of genomes from different organisms provides us with unprecedented opportunities to address many long-standing evolutionary questions in a more comprehensive way.

### 1.1 Lineage-specific genes

The availability of several genomes from related organisms permits the identification of newly evolved genes in different lineages or species, the study of their mechanisms of formation and the investigation of their role in adapting to new environments or physiological conditions (Domazet-Loso & Tautz, 2003; Guo et al., 2007; Khalturin et al., 2009; Kuo & Kissinger, 2008; Siepel, 2009; Toll-Riera et al., 2009a; Zhou et al., 2008). Recently formed genes give us the opportunity to study the action of natural selection in recent times and to investigate the processes associated with gene creation (Zhou & Wang, 2008).

The number of species-specific genes, or orphan genes, is not insignificant. They represent around 14% of the genes in 60 fully sequenced microbial genomes (Siew & Fischer, 2003) and between 20-30% in *Drosophila* species (Domazet-Loso & Tautz, 2003; Drosophila 12 Genomes Consortium, 2007). Genes restricted to particular lineages include vomeronasal receptors and casein milk proteins in mammals, which are known to be involved in specific physiological adaptations in this lineage (International Chicken Genome Sequencing Consortium, 2004). Additionally, several lineage-specific genes have been found to be involved in defence against pathogens, such as dermcidin in primates (Toll-Riera et al., 2009a) and surface antigens in apicomplexan parasites (Kuo & Kissinger, 2008). Interestingly, it has been noticed that rice orphan genes are more often expressed under environmental pressure (injury and hormone treatment) than non-orphan genes, indicating that novel genes help in adaptation to changing conditions (Guo et al., 2007).

Many newly evolved genes are derived from partial or complete gene duplication of pre-existing genes (Long et al., 2003; Marques et al., 2005; Toll-Riera et al., 2009a; Zhou et al., 2008). Alternative processes of gene formation include exaptation from mobile elements

(Nekrutenko & Li, 2001; Toll-Riera et al., 2009b), gene fusion or fission (Parra et al., 2006) and *de novo* gene formation from non-coding sequences (Cai et al., 2008; Heinen et al., 2009; Knowles & McLysaght, 2009; Levine et al., 2006; Toll-Riera et al., 2009a). A genome-wide study in *Drosophila melanogaster* has reported that gene duplication is the most common mechanism for the formation of novel genes in this species (Zhou et al., 2008).

## 1.2 Gene duplication

In the early thirties Haldane (Haldane, 1932) and Muller (Muller, 1935) were the first to propose gene duplication as a mechanism for the generation of new genes. Later, in the seventies, Ohno published an influential book about the role of gene duplication in evolution (Ohno, 1970), in which he emphasised the importance of gene duplication in generating protein functional diversity. With the availability of complete genome sequences it has become possible to estimate genomic rates of gene duplication (Lynch & Conery, 2000), analyse the pattern of evolution of the two duplicated copies, and identify lineage-specific gene family expansions. Expanded gene families that have been analysed in detail include olfactory receptors in mouse (Mouse Genome Sequencing Consortium 2002) and human (Gilad et al., 2005), and KRAB-associated zinc-finger in primates (Castresana et al., 2004). Genomic studies have shown that gene duplication is associated with increased coding sequence evolutionary rates (Lynch and Conery 2000; Scannell and Wolfe 2008), higher tissue expression divergence (Gu et al., 2002; Makova & Li, 2003), and higher regulatory sequence divergence (Farre & Alba, 2010).

The molecular mechanisms that have been proposed to be involved in duplication are non-allelic homologous recombination, transposon-mediated transposition and illegitimate recombination. The first two mechanisms imply the presence of sequence homology (Zhou et al., 2008). Yang and colleagues (Yang et al., 2008) found an excess of repetitive sequences at the breakpoints of the duplicated regions of a group of *Drosophila* lineage-specific young duplicates, suggesting the action of non-allelic homologous recombination. Another study in *Drosophila* found that dispersed duplicates have mainly arisen through non-allelic homologous recombination, while tandem duplicates most often arose through illegitimate recombination (Zhou et al., 2008). It has also been hypothesized that segmental duplications may arise from the recombination of Alu repeat sequences (Bailey et al., 2003).

Duplicated genes appear at a very high rate. It has been estimated that, on average, 0.01 duplicates arise per gene per million years (Lynch & Conery, 2000). The most frequent fate following gene duplication is believed to be the silencing of one of the duplicated copies due to the accumulation of degenerative mutations, a process that may take approximately 4 million years to complete (Lynch & Conery, 2000). However, sometimes both copies survive. The duplicated copy can acquire beneficial mutations and consequently gain a novel function with respect to the parental gene (neofunctionalisation), while the parental copy preserves its original function (Ohno, 1970). The duplicated copy may also be retained due to the split of the original function between the two gene copies (subfunctionalisation) (Hughes, 1994). Finally, if an increase in dosage of a particular gene is beneficial, the new copy may become fixed by positive selection maintaining the same gene structure and function as the parental gene (Kondrashov & Koonin, 2004).

Duplicated genes may confer adaptive advantages. For example, trichromatic colour vision in Old World Monkeys is associated with a pigment gene duplication that occurred after the



separation of New World Monkeys, and which gave rise to differentiated red and green pigments (Nathans et al., 1986). Zhang and colleagues (Zhang et al., 1998) reported on another example of the action of positive selection after gene duplication. The eosinophil cationic protein (ECP) and eosinophil-derived neurotoxin (EDN) genes are present in Old World Monkeys and hominoids, and probably originated by tandem gene duplication after the divergence of New World Monkeys. EDN is an antiviral agent (Domachowske & Rosenberg, 1997) and ECP is a potent toxin for bacteria and parasites (Rosenberg & Dyer, 1995). The authors detected a non-random accumulation of arginine substitutions in ECP, which may contribute to the generation of pores in pathogens' membranes. Another example refers to pancreatic ribonuclease 1B (RNASE1B), which originated through gene duplication of RNASE1, an enzyme used to digest bacteria in the small intestine, in the douc langur (*Pygathrix nemaeus*) around 2-4 million years ago (Zhang et al., 2002). Douc langurs are folivorous monkeys, in which leaves are digested through fermentation by symbiotic bacteria residing in the foregut. The newly duplicated copy, RNASE1B has evolved very rapidly (non-synonymous to synonymous nucleotide substitution rate of 4.03), contrary to the paralogous copy, RNASE1, which has not undergone change. These results indicate a burst of positive selection acting on the duplicated copy. Moreover, most of the substitutions imply the gain of negatively charged residues, lowering the optimal pH for RNASE1B, which could be related to an increase in digestive efficiency, given the lower pH found in the small intestine of douc langurs.

### 1.3 Partial gene duplication

Not all duplicated proteins are identical to their parental copies at birth. In fact, it has been reported that in *C. elegans* only about 40% of the new duplicates are borne out of complete gene duplications, the remainder representing cases of partial gene duplication (Katju & Lynch, 2003). These partially duplicated genes may recruit sequences from their genomic neighbourhood or from other genes (Katju & Lynch, 2006). In the first case, adjacent non-coding sequences are co-opted for a coding function. Katju and Lynch (Katju & Lynch, 2006) found that about half of the partially duplicated genes did not recruit any surrounding sequences but accumulated mutations, for example in initiation or termination codons, that altered the coding sequence. In *Drosophila melanogaster*, around 30% of the newly formed genes recruited various genomic sequences or formed chimeric gene structures (Zhou et al., 2008). Partially duplicated and chimeric genes are expected to adopt new functions immediately, which may increase their probability of being retained (Patthy, 1999; Zhou et al., 2008). An example of a gene that has arisen by partial duplication is the *Hun* gene in *Drosophila*, located on the X-chromosome. *Hun* arose from a partial duplication of the *Bällchen* gene, which is on chromosome 3R. *Hun* lacks 3' coding sequence with respect to *Bällchen*, but has gained 33 amino acids from a nearby intergenic sequence. Further, while *Bällchen* is expressed ubiquitously, *Hun* shows testes-specific expression (Arguello et al., 2006).

The sequence similarity that exists between completely duplicated gene copies and parental gene copies is often sufficient to detect homologues in a whole range of organisms. However, this is often not the case for partially duplicated genes, especially if the sequence common to both duplicates is short and the rate of divergence of the novel gene duplicate is abnormally high. As a result, many partially duplicated genes are identified as orphan or lineage-specific genes, that is, genes that do not yield any significant hits in database protein

searches of more distant organisms (Chen et al., 2010; Domazet-Loso & Tautz, 2003; Toll-Riera et al., 2009a). In a recent study that showed that newly formed genes in *Drosophila melanogaster* are as likely to perform essential functions as older genes, it was found that 28 out of the 50 new genes that had arisen through gene duplication corresponded to partial duplications (Chen et al., 2010). These young genes were found to evolve very rapidly, showing a median of 47.3% divergence, at the amino-acid level, from their parents. In an analysis of the mechanisms of formation of primate-specific genes, we observed that about 24% of the newly formed genes had originated through gene duplication, frequently involving partial gene duplication and the recruitment of additional sequences (Toll-Riera et al., 2009a). One example is human XAGE-1, a cancer/testis-associated gene that has partial homology to human XAGE-2, a gene that is well conserved in other mammals. The similarity is limited to the C-terminal half of the orphan XAGE-1 protein. We showed that, in the conserved region, the rate of amino acid sequence evolution of XAGE-1 was double that of XAGE-2, suggesting that the recruitment of additional sequences in XAGE-1 resulted in a marked asymmetry in the evolutionary rates of the two copies.

Partial gene duplication is likely to be very important for the formation of novel gene structures and the evolution of new protein functions, but studies focusing on this type of gene duplication are still scarce. To shed new light on this issue, we decided to analyse the evolutionary patterns of several primate-specific genes (orphan genes) formed, at least partially, by gene duplication. The results show that increased evolutionary rates in the partially duplicated copy are the norm, reinforcing the role of partial gene duplication in the formation of novel genes with distinct functions.

## 2. Results

Here we use a similar approach to that employed in Toll-Riera et al. (Toll-Riera et al., 2009a) to identify a set of primate-specific genes that show significant similarity to human genes (parental genes) that are well conserved in non-primate species. We investigate the differences in the rate of evolution of the novel and parental genes and discuss the role of partial duplication in increasing the protein functional repertoire.

### 2.1 Identification of primate lineage-specific genes formed by gene duplication

We identified a set of genes present in human and macaque but absent in 13 non-primate genomes (*Mus musculus*, *Rattus norvegicus*, *Bos Taurus*, *Canis familiaris*, *Gallus gallus*, *Xenopus tropicalis*, *Danio rerio*, *Takifugu rubripes*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Arabidopsis thaliana*). The existence of a homologue in a specific genome was determined by the presence of a BLASTP (Altschul et al., 1997) hit with an expectation value (E-value) smaller than  $10^{-4}$ , as previously described (Alba and Castresana 2005). Orphan genes were defined as those for which we could not detect any homologues in any of the species mentioned above. As they were, by definition, present in human and macaque, our collection of orphan genes corresponded to primate-specific genes, presumably formed after the split of the rodent and primate branches and before the speciation of the human and macaque lineages. Once we had this set of orphan genes, we investigated which ones could have arisen through gene duplication by performing BLASTP searches against all human proteins, using a relaxed E-value ( $E < 0.5$ ). We kept those cases for which we could identify human

paralogues that were not primate-specific. In such cases, the closest hit in human was considered the putative human parental gene, and the closest non-primate orthologue of the parental gene was taken as the outgroup gene (Figure 1). Protein sequences were aligned with T-Coffee (Notredame et al., 2000), and the alignments between primate-specific genes and parental genes were carefully examined to discard any spurious associations. We also removed any regions that were completely divergent (non-alignable) between the orphan and parental genes.

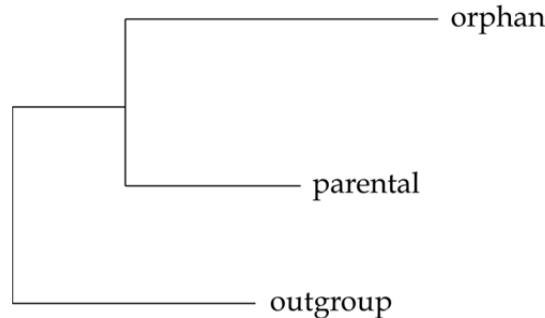


Fig. 1. Tree topology corresponding to gene families containing duplicated orphan genes. The orphan and parental genes are from human, the outgroup gene from a non-primate species.

The final set consisted of 14 orphan genes. Table 1 shows the orphan, parental and outgroup gene names, protein identifiers, and the percent of parental protein that could be reliably aligned with the orphan protein, corresponding to the portion of the protein that had duplicated. Of the 14 orphan genes, 4 represented single copies and the rest belonged to orphan gene families. In only one case, dermcidin, sequence similarity supported a complete gene duplication event.

We used the protein multiple alignments to estimate the number of amino acid substitutions per site (K) in the orphan, parental and outgroup branches. We used PROML, a maximum likelihood based method in the Phylip package for this purpose (Felsenstein, 2005). The results of these computations are discussed below.

## 2.2 Dermcidin and lacritin

The first example of an orphan gene that arose through gene duplication is dermcidin. This gene encodes a short protein of 110 amino acids in length. The corresponding parental gene is lacritin, which has orthologues in other mammals, and is located on chromosome 12 adjacent to the dermcidin gene. The two genes have a similar exonic structure, and although they are highly divergent, sequence similarity between the two is still detectable (Wang et al., 2006). Dermcidin is secreted in sweat glands, having an antimicrobial activity (Schitteck et al., 2001), and may also be involved in neural survival and cancer (Porter et al., 2003), whereas lacritin is expressed in the lacrimal glands (Ma et al., 2008).

Figure 2 shows the alignment of the complete protein sequences of human dermcidin, human lacritin and cat lacritin. The number of amino acid substitutions per site in the orphan branch was 1.026, about double the number of amino acid substitutions per site in the parental and outgroup branches (0.434 and 0.505, respectively).

Orphan Name	Parental name	Parental protein	Outgroup	%
Dermcidin	Lacritin	ENSP00000257867	ENSFCAP00000009317 (cat)	100%
FAM9A	Synaptonemal complex protein 3	ENSP00000266743	ENSMUSP000000020252 (mouse)	87.29%
FAM9B	idem	idem	idem	idem
FAM9C	idem	idem	idem	idem
AL023807.2	AL365202.1	ENSP00000382846	ENSCINP000000011125 (vase tunicate)	64.42%
XAGE-1A	XAGE-2	ENSP00000333775	XP_001249434.1 (cow)	36.03%
XAGE-1B	idem	idem	idem	idem
XAGE-1C	idem	idem	idem	idem
XAGE-1D	idem	idem	idem	idem
XAGE-1E	idem	idem	idem	idem
NP1P-like 1	Acyl-CoA synthetase medium-chain family member 1	ENSP00000428098	ENSMUSP000000036140 (mouse)	18,07%
C2orf27A	Ral guanine nucleotide dissociation stimulator like-4	ENSP00000290691	ENSCAFP000000031102 (dog)	12.05%
C2orf27B	idem	idem	idem	idem
AL133216.1	Arsenite-resistance protein 2	ENSP00000314491	ENSMUSP000000043123 (mouse)	9.36%

Table 1. List of primate-specific genes that have arisen by gene duplication. Protein identifiers are from Ensembl (ENSP) or Genbank (XP). % refers to the percentage of the parental protein that showed homology to the orphan protein.

### 2.3 Partially duplicated orphan genes

The remaining primate-specific genes that have arisen through gene duplication corresponded to partial duplications of the parental gene (Table 1). They included 3 individual genes (AL023807.2, NP1P-like 1 and AL133216.1) and 3 gene families (FAM9, XAGE-1 and C2orf27). The percentage of protein sequence from the parental protein that could be identified as homologous in the orphan protein ranged from 9.4 to 87.3% (Table 1). With the exception of NP1P-like 1, the orphan gene is located on a different chromosome from the parental gene, although the presence of introns in all orphan genes suggests that they were not retrotransposed copies. We aligned the conserved regions of orphan, parental and outgroup proteins (Figure 3). These alignments were used for the estimation of the

number of amino acid substitutions per site in the orphan, parental and outgroup branches. We also investigated the presence of any known protein domains in the region conserved between parental and orphan proteins, using the Pfam web server (Finn et al., 2010).

#### Dermcidin (ENSP00000293371)

```

Orphan   MRFMTLLFLTALAGALVCAYPEAASAFGSGNFCHEAS-----
Parental MKFTLLFLAAVAGALVYAEDAS--SDSTGADFAQEAGTSKFNEEISGPAEPASFPETTT
Outgroup MRFSALLLLAALAGALVCAQDAP--SDPTEATFGTVAETE--P-EVITSFPAETVFPFQET--

Orphan   AQKENAGED-----PGLARCAPKFRK---QRSS-LEKGLDGAKKAVGSLG-KLGKDAV
Parental TAQETSAAAVQGTAKVTSSRQELNPLKSIVEKSILLTEQALAKAGKGMHGGV-PGCKQFI
Outgroup -PQEPNSA-----TSKEGLNPLKLLVSKGSLVAEQQGFQEARKKLREGCKFERGVEELA

Orphan   EDLESVGGKGAVHDVKDVLDSVL
Parental ENGSEFAQKLLLKKF-SLLKPWA
Outgroup EKLKKFA-----PSFLLSV

```

Fig. 2. Alignment of dermcidin (orphan), human lacritin (parental) and cat lacritin (outgroup) proteins. Identical residues are in green, similar residues in yellow.

The FAM9 family (family with sequence similarity 9) is composed of three genes: FAM9A, FAM9B and FAM9C. They are all predicted to have a Cor1/Xlr/Xmr domain in the region of similarity to the parental gene (E-values ranging from 0.049 to  $4.4 \times 10^{-13}$ ), related to meiotic prophase chromosomes. The parental gene, synaptonemal complex protein 3 (SYCP3) is involved in the assembly of the synaptonemal complex during meiosis (Martinez-Garay et al., 2002), but the exact physiological functions of the FAM9 proteins remain unknown.

The largest orphan gene family is XAGE-1, which has 5 members with identical amino acid sequences that are contiguous on the X chromosome. The region conserved between the XAGE-1s and XAGE-2 includes the GAGE domain. The function of GAGE (G antigen) and XAGE (X antigen) domains is unknown, but XAGE and GAGE proteins have been implicated in several human cancers (Zendman et al., 2002).

The two genes belonging to the C2orf27 family are contiguous in the genome, though C2orf27A is located on the forward strand of chromosome 2 whereas C2orf27B is located on the reverse strand. Their function is unknown, but they derive from a protein annotated as Ral guanine nucleotide dissociation stimulator-like 4. The parental protein contains the RasGEF domain, which is a guanine nucleotide exchange factor for Ras-like small GTPases. The duplicated region overlaps minimally with this domain (14 amino acids).

NPPI-like 1 belongs to the nuclear pore complex-interacting protein (NPPI) family. The parental protein contains two AMP-binding domains that are at the N-terminal region of the protein, not the area conserved in the orphan protein, which is the C-terminal part. The NPPI family (Nuclear Pore Interacting Protein), also named *morpheus*, is located on a duplicated segment of chromosome 16. It has been suggested to have experienced a burst of positive selection during the emergence of *Homininae* (Johnson et al., 2001).

Finally, AL133216.1 and AL023807.2 are two primate-specific genes of unknown function containing putative coding sequences of length 151 and 121 amino acids respectively. The parental copy of AL133216.1 modulates arsenic sensitivity, is involved in cell cycle progression, and in RNA-mediated gene silencing by microRNA (Gruber et al., 2009). It also contains an arsenite-resistance protein 2 domain (Pfam hit E-value =  $3.1 \times 10^{-18}$ ). The orphan

copy does not contain this domain even though it is located in the conserved region, suggesting that this region has lost its ancestral function in the orphan protein.

Table 2 shows the estimated amino acid substitution rates in the orphan, parental and outgroup branches. In the case of identical copies (for example C2orf27A and C2orf27B) only one is taken as representative. In the case of divergent copies (the FAM9 family) the amino acid substitution rates are summed up for all branches from the ancestor to the derived node (see Figure 4). In all cases the duplicated protein is evolving much faster than the parental gene, and in some cases, such as the FAM9 and NPIP-like 1 proteins, more than six times faster. These results indicate that orphan proteins are evolving under much more relaxed constraints, and/or adapting to a new function with respect to their parental copies.

Orphan Name	Orphan protein	Orphan	Parental	Outgroup
Dermcidin	ENSP00000293371	1.02595	0.43458	0.5055
FAM9A	ENSP00000370391	1.28971	0.17014	0.15423
FAM9B	ENSP00000318716	1.13565	0.17014	0.15423
FAM9C	ENSP00000369999	1.15328	0.17014	0.15423
AL023807.2	ENSP00000381423	0.19840	0.12203	0.23096
XAGE-1A	ENSP00000382698	0.52961	0.17820	0.94188
NPIP-like 1	ENSP00000350444	0.40089	0.02020	0.11929
C2orf27B	ENSP00000304065	0.55865	0.21080	0.68342
AL133216.1	ENSP00000382606	1.37580	0.00010	0.00010

Table 2. Estimated number of amino acid substitutions per site (K) for orphan, parental and outgroup branches. Orphan protein identifiers are from Ensembl. See Table 1 for more details.

## 2.4 Role of low-complexity sequences

Low complexity regions (LCRs) are sequences in which one or a few residues are highly overrepresented. Several studies have shown that duplicated gene copies can gain new functions through the acquisition of LCRs (Fondon & Garner, 2004; Salichs et al., 2009). It has also been shown that young proteins contain more LCRs than old proteins (Alba & Castresana, 2005). Therefore, we inspected the presence of LCRs in our set of orphan proteins using the SEG algorithm with default parameters (Wootton & Federhen, 1996).

We found that the FAM9A protein contained a very conspicuous low-complexity sequence. Figure 4 shows the detailed phylogenetic tree of the FAM9 gene family (including the parental and outgroup SYCP3 genes). The ancestral FAM9 evolved very rapidly and eventually underwent two duplication events, leading to FAM9A, FAM9B and FAM9C. The multiple alignment of the region surrounding the LCR in FAM9A shows how, from a small region containing several acidic residues in SYCP3, a larger acidic region was formed in the common FAM9 ancestor, which finally expanded to a 75 amino acid stretch in FAM9A containing a long glutamic acid repeat, as well as poly-alanine and poly-glycine repeats.

As is the case for the SYCP3 proteins, all three human FAM9 proteins show testis-specific expression. However, the cellular localization is different depending on the protein studied: FAM9B and FAM9C are localized in the nucleus with low protein levels being detectable in the cytoplasm, whereas FAM9A is present at high levels in the nucleolus (Martinez-Garay et al., 2002). The distinct location of FAM9A may be due to the long glutamic acid repeat, as

FAM 9 (ENSP00000370391, ENSP00000318716, ENSP00000369999)

Orphan-FAM9B MAANGKK  
Orphan-FAM9A MEPVGRKRSRRAAKAQLAQTATGATKEGSGIASNFPQGPTMEPVGRKRSRRAAKAQL  
Orphan-FAM9C MA-AKDQLEVMMAA  
Parental MVSSGKKYSRKSGKSPVEDQFTRAY  
Outgroup MVPGGKRKSGKSGKPPVLDQPKAF

```

Orphan-FAM9B      -----HAG-GDVPVRDECEERNRFTETREEDVDTDEH-GEREPFAETDEHTGANTK
Orphan-FAM9A      EAQVRAAPAKKHTG-GDVPVRDECEERNFPTEREEDVDTDEH-GEREPFAEKDEHTGIHTM
Orphan-FAM9C      -----QEMELAG-GDVPVSHEEKERPVTETKEGDVDTDEH-GERGSFAETDEHTGVDTK
Parental          -----DFETEDKLDL-SGSEE-DVIEGA-TAVIKRRKKSGAGSVGDEDMGEVQN
Outgroup          -----DFEKDD-LL-SGSEE-DVIEGA-APVIKHGKKRSAGIEVDMGEVQN

```

Orphan-FAM9B KPE<sup>+</sup>DTA<sup>+</sup>---ED<sup>+</sup>TA<sup>+</sup>RRK<sup>+</sup>RM<sup>+</sup>KD<sup>+</sup>-KTC<sup>+</sup>-SKTKN<sup>+</sup>-KSKHA<sup>+</sup>RRKK<sup>+</sup>OLRQ<sup>+</sup>KRDY<sup>+</sup>IHS<sup>+</sup>LKLLN<sup>+</sup>

Orphan-FAM9A KLE<sup>+</sup>HI<sup>+</sup>AAD<sup>+</sup>IKG<sup>+</sup>LA<sup>+</sup>AK<sup>+</sup>EM<sup>+</sup>IK<sup>+</sup>D<sup>+</sup>KA<sup>+</sup>RYRTK<sup>+</sup>N-T<sup>+</sup>ERAL<sup>+</sup>KKK<sup>+</sup>OLRQ<sup>+</sup>KRDY<sup>+</sup>RH<sup>+</sup>TR<sup>+</sup>KL<sup>+</sup>LN<sup>+</sup>

Orphan-FAM9C ELE<sup>+</sup>DI<sup>+</sup>AAD<sup>+</sup>IKH<sup>+</sup>LA<sup>+</sup>AK<sup>+</sup>RR<sup>+</sup>IK<sup>+</sup>IA<sup>+</sup>KA<sup>+</sup>-SEIKN<sup>+</sup>-R<sup>+</sup>KNV<sup>+</sup>LT<sup>+</sup>OLRQ<sup>+</sup>KRDY<sup>+</sup>RH<sup>+</sup>TR<sup>+</sup>KL<sup>+</sup>LN<sup>+</sup>

Parental MLE<sup>+</sup>GV<sup>+</sup>GD<sup>+</sup>INKA<sup>+</sup>LAK<sup>+</sup>RR<sup>+</sup>KLE<sup>+</sup>MYT<sup>+</sup>KAS<sup>+</sup>-LKTS<sup>+</sup>NK<sup>+</sup>IE<sup>+</sup>HW<sup>+</sup>VK<sup>+</sup>TD<sup>+</sup>QER<sup>+</sup>OLK<sup>+</sup>LNQ<sup>+</sup>YS<sup>+</sup>QPF<sup>+</sup>MT<sup>+</sup>

Outgroup MLE<sup>+</sup>KFG<sup>+</sup>VD<sup>+</sup>INKA<sup>+</sup>LAK<sup>+</sup>RR<sup>+</sup>IKLE<sup>+</sup>MYT<sup>+</sup>KAS<sup>+</sup>-FKASH<sup>+</sup>NK<sup>+</sup>IE<sup>+</sup>HW<sup>+</sup>VK<sup>+</sup>TD<sup>+</sup>QER<sup>+</sup>OLK<sup>+</sup>LNQ<sup>+</sup>YS<sup>+</sup>QPF<sup>+</sup>MT<sup>+</sup>

Orphan-FAM9B VLE E Y T D E Q K E E  
Orphan-FAM9A VI K E Y A E K Q D D A E E A E A A A A A E A A A A A E A A A A A E V I V V D E E E E E E E E E E E E E E E  
Orphan-FAM9C V L E E F T D E Q K D E E  
Parental LFQ Q W D L D M Q K A E E  
Outgroup V L Q O W E L D I Q K F E E

```

Orphan-FAM9B  -----EEEGEEEEELIRIQEQQKNQOYKSVRRE
Orphan-FAM9A  EEEGEEEGGGGEEGGGGGGGEGEETEEEEEEEEE  EEEEEEQIKAFQEKRRNQDPTGVRSSW
Orphan-FAM9C  -----GDGKEEQIKAFQEKRRNQDQGGKTER
Parental      -----QEEKILNMRQQQKILQDSIVQSQ
Outgroup      -----QGEKLSNMRQQQKIFQDSIVQSQ

```

Orphan-FAM9B **RL**KEMK**LL**R**DQ**FV**K**A**LE**D**F**ED**LC**R**V**F**S**DE**D**SE**LD**N  
Orphan-FAM9A **RL**RE**M**K**PL**LE**Q**LL**K**A**K**D**TKD****N**Y**C**I**S**SE**E**SE**LD**N  
Orphan-FAM9C D-  
Parental **RL****K**T**I****Q****L****Y****E****Q****F****I****K****S****M****E****E****L****K****N****H****D****N****L****L****T****G****A****Q****N****E****F****K****K**  
Outgroup **RM****K****A****I****K****O****I****H****E****O****F****I****K****S****L****E****D****V****E****K****N****N****L****F****T****G****T****S****E****L****K****K**

AL023807.2 (ENSP00000381423)

Orphan **FFIFLPRQSLALLPRLECSGTISAHCNLSLPGSSISHASAFRVAGITGVHCRITQLIFVYLVETGPFHMM**  
 Parental **FFFLPKKSLALLPRLEYSGTILAHCNLCILGSSNSHASASQVAGITGMCHHVVMLIFVFLVETGPFHVV**  
 Outgroup **FFFLRWSITLLPRLECSGAISHAHCNRLPGSSDPTTASQVAGITGMCHHSOLIFVLVDLMRFHCV**

XAGE-1A (ENSP00000382698)

Orphan KSCISQTPGINLDLGSQVVKVILPKKEHC KMPEAGEEQPQV  
Parental ELCQTKT-GDGCEGGTDVKGKILPKAEHFKMPEAGEGKSQV  
Outgroup OLAVAKTGGEG-GDGPDVREEFASNTEPGEMPEAGEGQFFA

NPIP-like 1 (ENSP00000350444)

Orphan **Q**MV**K**LSIVLTPQFLSHD**Q**S**S**FT**K**ELQ**Q**HVKS**V**TC**P**CEY**L**RKV**I**NS  
 Parental EVVKAFIVLTPQFLSHD**K**D**Q**L**T**KELQ**Q**HVKS**V**TAPYKYPRKVEF**V**  
 Outgroup EVVKAFIVL**N**PEFLSHD**E****O**L**I**KELQ**H**HVKS**V**TAPYKYPRKVEF**V**

C2orf27B (ENSP00000304065)

Orphan **PVPAPGADSF**-PGTALELEEAPEPSRCPTAQDQPSEELPDFMAPPPVEPPASALELK  
Parental PAPAPGEGFP-PGTVLEPQSAPESSCPCRGSVKNQPSSELPDMTTFPPRLLAEQLTLM  
Outgroup PAFV-GPDTFSLSTKTEPPAPEATCH-WGTPAHORHEEOGLMAFPKKLVAEOLTSI

AL133216.1 (ENSP00000382606)

Orphan PQLPEAKLPRPSSGLTVASPGSAPALRWHLQAPNGLRSVGSRRPS-LGLPAASAGFKRFE  
Parental PALPEIKPAQPPGPAQIILPGLTLPGLPYPHQTPOGLMPYGOFRPPIILGYGAGAVRPAVPT  
Outgroup PALPEIKPAQPPGPAOILPGLTLPGLPYPHQTPOGLMPYGOFRPPIILGYGAGAVRPAVPT

Orphan VGLSRPSS-----GLLAAFAG  
Parental GGPPYPHAPYGAGRGNYDAFRG  
Outgroup GGPPYPHAPYGAGRGNYDAFRG

Fig. 3. Multiple alignments of the conserved regions between orphan, parental and outgroup proteins. For the XAGE-1 family only XAGE-1A is shown, as the other orphan sequences were identical at the amino acid level. The same is true for the C2orf27 family. Identical residues are in green, similar residues in yellow. See Table 1 for more details.

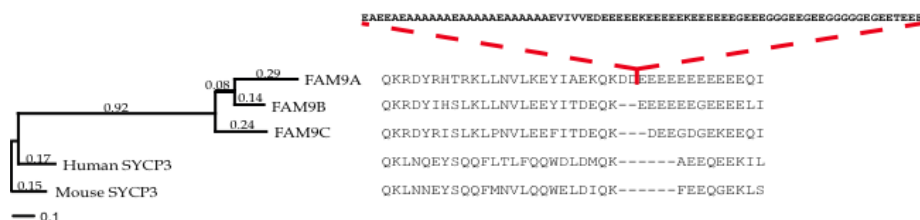


Fig. 4. Phylogenetic tree of the FAM9 gene family. Branch lengths correspond to the estimated number of amino acid substitutions per site, using the alignment in Fig. 3. The protein alignment shown corresponds to exon 5 in FAM9B and FAM9C and to exon 6 in FAM9A, human SYCP3 and mouse SYCP3. The expanded low-complexity region in FAM9A is depicted above the alignment.

acidic clusters have been shown to mediate protein nucleolar retention (Ochs et al., 1996; Shu-Nu et al., 2000; Ueki et al., 1998). In FAM9A, the low complexity sequence is located within the Cor1/Xlr/Xmr conserved region, perhaps interfering with its function. In fact, FAM9A shows higher sequence divergence from the common ancestor than FAM9B.

### 3. Discussion

The role of partial gene duplication in the formation of novel genes is still poorly understood, although recent reports in *Drosophila* (Chen et al., 2010; Zhou et al., 2008) and *C.elegans* (Katju & Lynch, 2006; 2003) indicate that partially duplicated gene copies are very frequent. The present study analyses a set of primate-specific genes formed by partial gene duplication. We find that the rate of divergence of the partially duplicated copy is, in all cases, higher than the rate of divergence of the parental copy, generalizing previous observations for XAGE1-A (Toll-Riera et al., 2009a). This, together with the fact that most partially duplicated genes recruit additional sequences, strengthens the notion that partial duplication is a major process for the formation of genes with novel structures and functions. In these genes, any remaining similarity to the homologous proteins is being quickly erased by high sequence turnover. As a consequence, distant homologues are difficult to identify and these proteins end up being classified as orphans. This fits the model of Domazet-Loso and Tautz in explaining the high number of orphan genes in *Drosophila*: orphan genes are created by gene duplication followed by a period of rapid sequence divergence that erases the similarity with its homologues (Domazet-Loso & Tautz, 2003). Although we now have evidence that not all orphan genes are generated in this manner (Toll-Riera et al., 2009a; Toll-Riera et al., 2009b; Zhou et al., 2008), a significant portion is.

A large fraction of the duplicated gene copies that become fixed in a population are subsequently lost, presumably because the new copy is completely redundant and thus dispensable. However, the formation of chimeric gene structures, encoding part of an existing protein together with additional sequences, could in principle favour their retention, as these genes are not going to be functionally equivalent to the ancestral gene (Patthy, 1999; Zhou et al., 2008). In support of this, in *Drosophila* it was found that the proportion of novel genes corresponding to complete gene duplications decreased with gene age, suggesting that complete gene duplications had a shorter lifespan than partial gene duplications (Zhou et al., 2008).



Orphan genes are in general poorly annotated and their function is unknown in most cases (Kuo & Kissinger, 2008). The fact that organisms had lived perfectly well without them until recent times when they made their appearance, has led scientists to think that orphan genes were, for the most part, dispensable. However, a recent study by Chen and colleagues (Chen et al., 2010) has challenged this viewpoint. In their study, the authors identified new young genes in *Drosophila melanogaster* (around 34 million years old) and designed RNA interference lines to knock each of them out (KO). Surprisingly, they found that 30% of these young genes KOs were lethal, as *Drosophila* could not survive without them. These young genes had mainly arisen through duplication and they showed higher evolutionary rates than the parental gene, indicating the action of positive selection, or relaxation of functional constraints. They hypothesized that new genes are quickly integrated into existing pathways, and hence many of them soon become essential for the viability of the organism.

Capra and colleagues (Capra et al., 2010) compared the evolutionary patterns of genes that arose by duplication with those that did not (named novel genes). They argued that the evolutionary pressures should be different in each case as, contrary to novel genes, duplicated genes were functionally and structurally well formed from birth. They showed that although duplicated genes are initially more integrated into cellular networks, both types of new genes gain functions and interactions with time, though novel genes do it more rapidly than duplicated genes. Additionally, novel genes also increase in length through the incorporation of transposable elements or surrounding sequences. This increase in length could be related with the rapid gain of function and interactions experienced by novel genes. They also found that genes tended to interact with genes similar in age and mode of origin. Thus, the mechanism by which a gene originates seems to significantly impact on its subsequent evolution.

Several studies have demonstrated that duplicated genes show increased protein evolutionary rates with respect to non-duplicated genes in the same lineage (Castillo-Davis et al., 2004; Cusack & Wolfe, 2007; Kondrashov et al., 2002; Lynch & Conery, 2000; Nembaware et al., 2002; Scannell & Wolfe, 2008; Van de Peer et al., 2001). Here we identified a very strong asymmetry in the rates of evolution of the newly evolved copy (orphan) and the well-conserved copy (parental), the former evolving much faster than the latter. Surprisingly, the parental protein copy did not evolve consistently faster than the outgroup protein (not duplicated), highlighting the fact that we are dealing with a special type of gene duplication in which the copy containing the partially duplicated segment rapidly departs from the ancestral family, which remains essentially unaffected.

Increased evolutionary rates may reflect either relaxation of purifying selection, positive selection, or the combined effects of both these forces. The orphan genes under study predated the split of the human and macaque lineages, which occurred approximately 25 million years ago so, if relaxed selection was the only factor for their increased rates, the genes should by now have become pseudogenes and not be expressed. However, all genes were expressed at the RNA level in one or several tissues. Therefore we must hypothesize that, at least to some extent, positive selection has influenced the evolution of these genes.

We compared the rates of evolution of the protein regions that were conserved between orphan and parental proteins, but what about the unique sequences contained in the orphan proteins? These sequences lacked any similarity to other protein-coding genes, so they may be ancestral non-coding sequences that have been co-opted for a coding function (Long et al., 2003). Genes generated *de novo* from non-coding sequences are among the fastest evolving genes (Levine et al., 2006), and there is no reason to believe that unique sequences

in orphan proteins will evolve slower than the conserved protein regions, rather the contrary would seem more logical. In a previous study we showed that the non-synonymous to synonymous nucleotide substitution rates of primate-specific genes, measured for human and macaque orthologues, were, on average, twice as high as those of mammalian-specific genes and five times higher than those of deeply conserved eukaryotic proteins (Toll-Riera et al., 2009a). The differences in amino acid substitution rates between orphan and parental genes described here reinforce the idea that the evolution of a new gene is strongly associated with very rapid sequence change.

#### 4. Concluding remarks and future research

We have examined the evolutionary dynamics of a group of novel primate-specific genes (orphan genes) that have arisen by gene duplication. These genes typically form new structures in which only part of the protein sequence is shared with the parental copy, presumably because of partial gene duplication, and the rest of the protein sequence is unique. The orphan proteins accumulate a much larger number of amino acid substitutions per site than the parental proteins, denoting rapid functional diversification. The parental gene copies appear to act as “donors” of sequence but do not experience any obvious sequence evolution alterations, thus they probably preserve their ancestral functions. Future research in this area, using computational as well as experimental studies, should help clarify how frequent is partial gene duplication with respect to complete gene duplication, the differences in gene copy survival in both cases, and how partial and complete gene duplication contribute to the generation of evolutionary novelties.

#### 5. Acknowledgments

We received financial support from Ministerio de Educación, Gobierno de España (FPU to M.T-R, BIO2009-08160), Generalitat de Catalunya (FI to S.L.), Fundació Javier Lamas Miralles (Ajut Predoctoral Javier Lamas Miralles to N.R-T) and Institució Catalana de Recerca i Estudis Avançats (M.M.A).

#### 6. References

- Alba, M.M. & Castresana, J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol Biol Evol* 22(3): 598-606.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17): 3389-3402.
- Arguello, J.R., Chen, Y., Yang, S., Wang, W. & Long, M. 2006. Origination of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila*. *PLoS Genet* 2(5): e77.
- Bailey, J.A., Liu, G. & Eichler, E.E. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* 73(4): 823-834.
- Cai, J., Zhao, R., Jiang, H. & Wang, W. 2008. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* 179(1): 487-496.
- Capra, J.A., Pollard, K.S. & Singh, M. 2010. Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biol* 11(12): R127.

- Castillo-Davis, C.I., Hartl, D.L. & Achaz, G. 2004. cis-Regulatory and protein evolution in orthologous and duplicate genes. *Genome Res* 14(8): 1530-1536.
- Castresana, J., Guigo, R. & Alba, M.M. 2004. Clustering of genes coding for DNA binding proteins in a region of atypical evolution of the human genome. *J Mol Evol* 59(1): 72-79.
- Chen, S., Zhang, Y.E. & Long, M. 2010. New genes in *Drosophila* quickly become essential. *Science* 330(6011): 1682-1685.
- Cusack, B.P. & Wolfe, K.H. 2007. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol Biol Evol* 24(3): 679-686.
- Domachowski, J.B. & Rosenberg, H.F. 1997. Eosinophils inhibit retroviral transduction of human target cells by a ribonuclease-dependent mechanism. *J Leukoc Biol* 62(3): 363-368.
- Domazet-Loso, T. & Tautz, D. 2003. An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res* 13(10): 2213-2219.
- Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450(7167): 203-218.
- Farre, D. & Alba, M.M. 2010. Heterogeneous patterns of gene-expression diversification in mammalian gene duplicates. *Mol Biol Evol* 27(2): 325-335.
- Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
- Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R. & Bateman, A. 2010. The Pfam protein families database. *Nucleic Acids Res* 38(Database issue): D211-222.
- Fondon, J.W., 3rd & Garner, H.R. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A* 101(52): 18058-18063.
- Gilad, Y., Man, O. & Glusman, G. 2005. A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res* 15(2): 224-230.
- Gruber, J.J., Zatechka, D.S., Sabin, L.R., Yong, J., Lum, J.J., Kong, M., Zong, W.X., Zhang, Z., Lau, C.K., Rawlings, J., Cherry, S., Ihle, J.N., Dreyfuss, G. & Thompson, C.B. 2009. Ars2 links the nuclear cap-binding complex to RNA interference and cell proliferation. *Cell* 138(2): 328-339.
- Gu, Z., Nicolae, D., Lu, H.H. & Li, W.H. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* 18(12): 609-613.
- Guo, W.J., Li, P., Ling, J. & Ye, S.P. 2007. Significant comparative characteristics between orphan and nonorphan genes in the rice (*Oryza sativa* L.) genome. *Comp Funct Genomics*: 21676.
- Haldane, J.B.S. 1932. The causes of evolution. London: Longmans and Green.
- Heinen, T.J., Staubach, F., Haming, D. & Tautz, D. 2009. Emergence of a new gene from an intergenic region. *Curr Biol* 19(18): 1527-1531.
- Hughes, A.L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* 256(1346): 119-124.
- International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432(7018): 695-716.

- Johnson, M.E., Viggiano, L., Bailey, J.A., Abdul-Rauf, M., Goodwin, G., Rocchi, M. & Eichler, E.E. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* 413(6855): 514-519.
- Katju, V. & Lynch, M. 2003. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. *Genetics* 165(4): 1793-1803.
- Katju, V. & Lynch, M. 2006. On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome. *Mol Biol Evol* 23(5): 1056-1067.
- Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T.C. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet* 25(9): 404-413.
- Knowles, D.G. & McLysaght, A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res* 19(10): 1752-1759.
- Kondrashov, F.A. & Koonin, E.V. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet* 20(7): 287-290.
- Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I. & Koonin, E.V. 2002. Selection in the evolution of gene duplications. *Genome Biol* 3(2): RESEARCH0008.
- Kuo, C.H. & Kissinger, J.C. 2008. Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. *BMC Evol Biol* 8: 108.
- Levine, M.T., Jones, C.D., Kern, A.D., Lindfors, H.A. & Begun, D.J. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A* 103(26): 9935-9939.
- Long, M., Betran, E., Thornton, K. & Wang, W. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* 4(11): 865-875.
- Lynch, M. & Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290(5494): 1151-1155.
- Ma, P., Wang, N., McKown, R.L., Raab, R.W. & Laurie, G.W. 2008. Focus on molecules: lacritin. *Exp Eye Res* 86(3): 457-458.
- Makova, K.D. & Li, W.H. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res* 13(7): 1638-1645.
- Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A. & Kaessmann, H. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol* 3(11): e357.
- Martinez-Garay, I., Jablonka, S., Sutajova, M., Steuernagel, P., Gal, A. & Kutsche, K. 2002. A new gene family (FAM9) of low-copy repeats in Xp22.3 expressed exclusively in testis: implications for recombinations in this region. *Genomics* 80(3): 259-267.
- Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915): 520-562.
- Muller, H.J. 1935. The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. *Genetica* 17: 237-252.
- Nathans, J., Thomas, D. & Hogness, D.S. 1986. Molecular genetics of human color vision: the genes encoding blue, green, and red pigments. *Science* 232(4747): 193-202.
- Nekrutenko, A. & Li, W.H. 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* 17(11): 619-621.

- Nembaware, V., Crum, K., Kelso, J. & Seoighe, C. 2002. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome Res* 12(9): 1370-1376.
- Notredame, C., Higgins, D.G. & Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302(1): 205-217.
- Ochs, R.L., Stein, T.W., Jr., Chan, E.K., Ruutu, M. & Tan, E.M. 1996. cDNA cloning and characterization of a novel nucleolar protein. *Mol Biol Cell* 7(7): 1015-1024.
- Ohno, S. 1970. Evolution by gene duplication. *New York: Springer-Verlag*.
- Parra, G., Reymond, A., Dabbouseh, N., Dermitzakis, E.T., Castelo, R., Thomson, T.M., Antonarakis, S.E. & Guigo, R. 2006. Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res* 16(1): 37-44.
- Patthy, L. 1999. Genome evolution and the evolution of exon-shuffling--a review. *Gene* 238(1): 103-114.
- Porter, D., Weremowicz, S., Chin, K., Seth, P., Keshaviah, A., Lahti-Domenici, J., Bae, Y.K., Monitto, C.L., Merlos-Suarez, A., Chan, J., Hulette, C.M., Richardson, A., Morton, C.C., Marks, J., Duyao, M., Hruban, R., Gabrielson, E., Gelman, R. & Polyak, K. 2003. A neural survival factor is a candidate oncogene in breast cancer. *Proc Natl Acad Sci U S A* 100(19): 10931-10936.
- Rosenberg, H.F. & Dyer, K.D. 1995. Eosinophil cationic protein and eosinophil-derived neurotoxin. Evolution of novel function in a primate ribonuclease gene family. *J Biol Chem* 270(37): 21539-21544.
- Salichs, E., Ledda, A., Mularoni, L., Alba, M.M. & de la Luna, S. 2009. Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. *PLoS Genet* 5(3): e1000397.
- Scannell, D.R. & Wolfe, K.H. 2008. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res* 18(1): 137-147.
- Schitteck, B., Hipfel, R., Sauer, B., Bauer, J., Kalbacher, H., Stevanovic, S., Schirle, M., Schroeder, K., Blin, N., Meier, F., Rassner, G. & Garbe, C. 2001. Dermcidin: a novel human antibiotic peptide secreted by sweat glands. *Nat Immunol* 2(12): 1133-1137.
- Shu-Nu, C., Lin, C.H. & Lin, A. 2000. An acidic amino acid cluster regulates the nucleolar localization and ribosome assembly of human ribosomal protein L22. *FEBS Lett* 484(1): 22-28.
- Siepel, A. 2009. Darwinian alchemy: Human genes from noncoding DNA. *Genome Res* 19(10): 1693-1695.
- Siew, N. & Fischer, D. 2003. Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins* 53(2): 241-251.
- Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X. & Alba, M.M. 2009a. Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol* 26(3): 603-612.
- Toll-Riera, M., Castelo, R., Bellora, N. & Alba, M.M. 2009b. Evolution of primate orphan proteins. *Biochem Soc Trans* 37(Pt 4): 778-782.
- Ueki, N., Kondo, M., Seki, N., Yano, K., Oda, T., Masuho, Y. & Muramatsu, M. 1998. NOLP: identification of a novel human nucleolar protein and determination of sequence requirements for its nucleolar localization. *Biochem Biophys Res Commun* 252(1): 97-102.

- Van de Peer, Y., Taylor, J.S., Braasch, I. & Meyer, A. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J Mol Evol* 53(4-5): 436-446.
- Wang, J., Wang, N., Xie, J., Walton, S.C., McKown, R.L., Raab, R.W., Ma, P., Beck, S.L., Coffman, G.L., Hussaini, I.M. & Laurie, G.W. 2006. Restricted epithelial proliferation by lacritin via PKC $\alpha$ -dependent NFAT and mTOR pathways. *J Cell Biol* 174(5): 689-700.
- Wootton, J.C. & Federhen, S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* 266: 554-571.
- Yang, S., Arguello, J.R., Li, X., Ding, Y., Zhou, Q., Chen, Y., Zhang, Y., Zhao, R., Brunet, F., Peng, L., Long, M. & Wang, W. 2008. Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila*. *PLoS Genet* 4(1): e3.
- Zendman, A.J., Van Kraats, A.A., Weidle, U.H., Ruiter, D.J. & Van Muijen, G.N. 2002. The XAGE family of cancer/testis-associated genes: alignment and expression profile in normal tissues, melanoma lesions and Ewing's sarcoma. *Int J Cancer* 99(3): 361-369.
- Zhang, J., Rosenberg, H.F. & Nei, M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci U S A* 95(7): 3708-3713.
- Zhang, J., Zhang, Y.P. & Rosenberg, H.F. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat Genet* 30(4): 411-415.
- Zhou, Q. & Wang, W. 2008. On the origin and evolution of new genes--a genomic and experimental perspective. *J Genet Genomics* 35(11): 639-648.
- Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S. & Wang, W. 2008. On the origin of new genes in *Drosophila*. *Genome Res* 18(9): 1446-1455.

## **Part 2**

### **A Look at Some Gene Families**





# Immunoglobulin Polygeny: An Evolutionary Perspective

J. E. Butler, Xiu-Zhu Sun and Nancy Wertz

*Department of Microbiology & Interdisciplinary Immunology Program  
Carver College of Medicine, University of Iowa, Iowa City,  
USA*

## 1. Introduction

The immune system of vertebrates is characterized by genes of the Ig-superfamily (IGSF) that encode the immunoglobulin (Ig) genes, genes that encode the T cell receptor (TCR), a portion of the structure of the genes encoding the major histocompatibility molecules (MHC), Ig cell surface and transport receptors, some families of cytokines and chemokines as well as numerous other proteins important to the immune system. IGSF genes also encode proteins in sponges, coelenterates and flatworms (Blumbach et al., 1998; Miller & Steele, 2000; Ogawa et al., 1998). While not a topic for this chapter, we acknowledge that the IGSF genes are not the only family of genes used to generate an antibody repertoire in vertebrates. The VLR-based receptors of jawless fishes that belong to the LRR family of receptors, have had a parallel evolution (Herrin & Cooper, 2010).

Figure 1 illustrates the signature features of proteins encoded by the IGSF genes. Highly diagnostic is the so-called “ $\beta$ -barrel” or “Ig fold”. Anti-parallel  $\beta$ -pleated sheets form the staves of the barrel that are joined at each end by flexible polypeptide chains. These flexible polypeptides on the face of a heavy chain variable region domain (VH; Fig. 1A) contain three combinatorial determining regions (CDRs). The variable light chain domain (VL; not shown) also contributes three CDRs. CDRs from both VH and VL domains coalesce to form the antibody binding site (Fig. 1B). A striking feature of IGSF genes that encode the variable region domain of Igs is the degree of polygeny such that duplicated VH genes alone can occupy > three megabases (Matsuda et al., 1990). There are three such variable region loci in mammals: VH, VL and V $\kappa$ . The former encodes the variable heavy chain domain (Fig. 1A) while VL and V $\kappa$  encode the light chains variable region domains. All three loci are independent (non-linked) although a few orphan human VH genes can be found in other linkage groups (Matsuda et al., 1990). Popular textbooks suggest that this polygeny explains why antibodies can recognize >10<sup>10</sup> different antigens. It is argued that if each specific antibody required a completely separate gene, more DNA would be needed than exists in the mammalian genome. To reduce the need for so many different germline encoded antibody binding sites, a system of somatic gene segment recombinations and later, somatic hypermutation (SHM) or somatic gene conversion (SCG), evolved.

The complete antibody molecule (and other proteins encoded by IGSF genes) is often composed of a tandem series of  $\beta$ -barrel domains as illustrated in Fig. 1B. Each domain in such multi-domain molecules differs slightly in structure and correspondingly, in function.

In this article we will focus on the genes encoding the VH domain of Ig (Fig. 1A) as well as those encoding the so-called “constant regions” (C $\gamma$ ) of IgG, the mammalian flagship antibody isotype. We use as examples the sequences of duplicated VH genes in swine and bats (opposite extremes) and the duplicated C $\gamma$  genes encoding the subclasses of swine IgG, as evidence to suggest how this polygeny occurred.

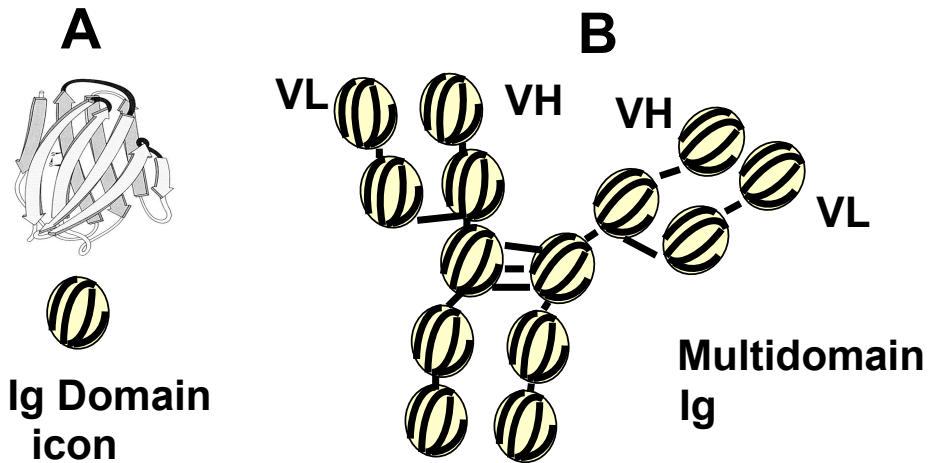


Fig. 1. Duplication/diversification of Ig genes resulted in macromolecules with repeating units. A. The variable heavy chain domain (VH) with its characteristic  $\beta$ -barrel or Ig fold. The dark polypeptides connecting the “barrel staves” contain the CDR regions. B. Complete Igs are multidomain molecules comprised of many Ig fold domains. The CDR-containing peptides occur only in the VH and VL domains. The remaining C-domains comprise the constant region of the Ig. The “monomeric Ig” shown in Fig. 1B is bivalent, with two identical VH/VL pairs that contain the antigen binding sites.

In the interests of those who are not immunologists, we describe the different Ig-loci, how they vary among vertebrates and the processes involved in the generation of the antibody repertoire (Section 2). Section 3 discusses the gene duplication phenomenon which resulted in the polygeny that characterizes the vertebrate Ig genome, while Section 4 reviews the somatic processes that lead to the synthesis and secretion of antibodies in higher vertebrates. Section 5 discusses the selection processes involved in gene usage. Finally, we provide data from studies in fetal/neonatal piglets, newborn rabbits and the chicken, to support the view that only a small number of the many duplicated V-region Ig genes are actually used. We provide examples in which only one or a few VH genes are needed to generate the antibody repertoire so long as the machinery for somatic recombination and somatic mutation is in place. Based on these examples and comparing them to antibody repertoire development in lower vertebrates, we hypothesize that the extensive polygeny in the Ig loci of higher vertebrates exists as an evolutionary vestige but is retained because of its redundancy value. While the recent duplication/diversification of C $\gamma$  allows for specialized effector function, IgG in rabbits did not diversify, yet few would argue against the success of this mammalian order. Thus, many of the C $\gamma$  duplicons in other mammals may also have been retained for

their redundancy value. This could explain why individuals with major deletions of certain  $C\gamma$  genes remain healthy (see Section 6.3).

## 2. Organization of the Ig Loci

### 2.1 Translocon organization of gene segments characterizes higher vertebrates

Figure 2A shows the organization of the light and heavy chain loci. Each locus can be divided into subloci that, from 5' to 3', are known as the V, D, J and C regions. The light chain loci are similar but lack D subloci. As discussed above, the V, D, and J regions encode the antibody binding site for the heavy and light chain, and are comprised of a large number of duplicated gene segments that vary among species (Table 1). The  $V_H$  and  $V_L$  gene segments are the largest ( $\sim 300$  nucleotides) and encode both framework regions (FR) and CDR1 and CDR2. The FR regions encode the  $\beta$ -pleated sequences of the  $\beta$ -barrel (Fig. 1A). Displayed in linear fashion FR1, CDR1, FR2, CDR2 and FR3 comprise a  $V_H$  (or  $V_L$ ) gene (Fig. 4 and 5). The 3' portion of the  $J_H$  segment (after the tryptophan codon; Fig. 8) encodes FR4 while CDR3 results from the recombination of V-D-J or V-J (see Section 4).

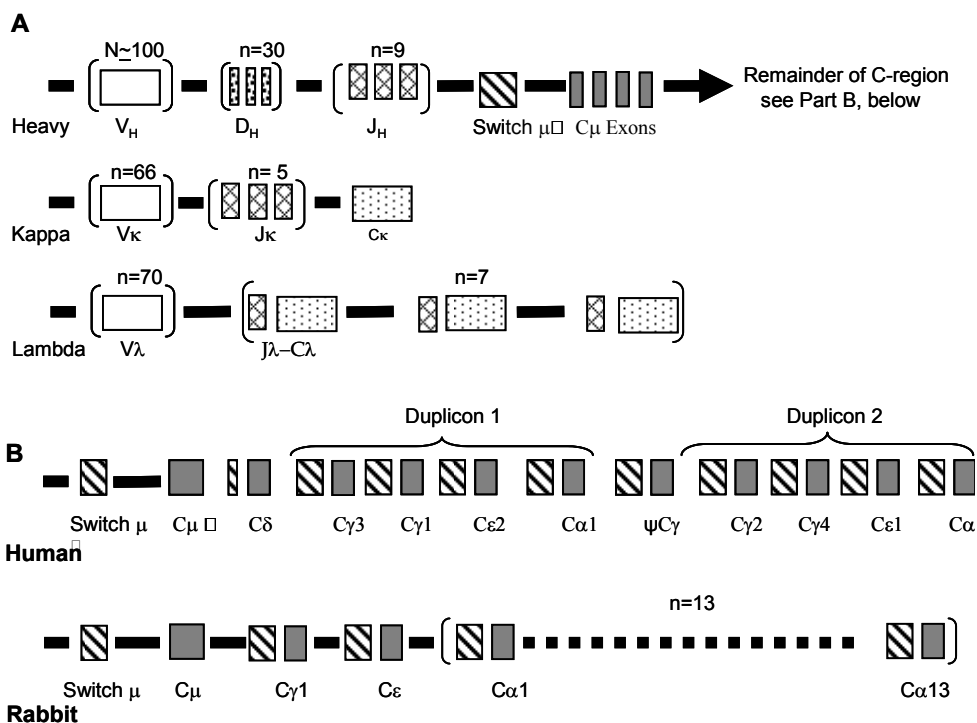


Fig. 2. The translocon organization of Ig genes of mammals. A. Organization of the variable region gene segments of the human heavy chain ( $V_H$ ), kappa ( $\kappa$ ) and lambda ( $\lambda$ ) loci. Brackets indicate the number (n) of gene segments of a particular type. Switch regions are depicted with diagonal strips. B. Organization of the constant region of the heavy chain locus of human and rabbit. The site of intralocus segment duplication in humans is indicated.

The C- region Sublocus is composed of exons of the genes that encode the “constant” domains of the antibody molecule (Fig. 1B). Fig. 2A illustrates the four exons that encode IgM (C $\mu$ ). Each exon encodes one of the constant region domains illustrated in Fig. 1B. Each domain possesses the  $\beta$ -barrel structures that characterizes the minimal structure of proteins encoded by IGSF genes (Fig. 1B). Within the C-region sublocus, sets of exons encode different antibody isotypes: e.g. IgM, IgD, IgG, IgE IgA (Fig. 2B). Two variations of the C-region sublocus are illustrated by the human and rabbit. The distribution of the encoded isotypes among common vertebrates is summarized in Table 2. Of interest to the theme of this review, the C-region sublocus contains a region that contains stretches of exons of duplicated C $\gamma$  genes encoding IgG subclasses (Fig. 2B; most mammals) or multiple IgA subclasses (Fig. 2B; in rabbit).

Species	V <sub>H</sub> (F*)	D <sub>H</sub>	J <sub>H</sub>	V <sub>L</sub> (F*)	J <sub>L</sub>	C <sub>L</sub>	V <sub>K</sub> (F*)	J <sub>K</sub>	C <sub>K</sub>
Human	87 (7)	30	9	70 (7)	7	7**	66 (7)	5	1
Mouse	>100 (14)	11	4	3 (3)	4	4**	140 (4)	4	1
Rat	>100 (11)	?	5	15 (4)	1	1	18 (?)	6	?
Rabbit	>100 (1)	12	6	?	2	2	>36 (?)	5	2
Swine	> 20 (1)	2	1	?	>3	>3	60 (2)	5	1
Horse	>10 (2)	>7	>5	25 (3)	4	4	>20 (?)	5	1
Cattle	>15 (2)	3	5	30 (?)	>2	4	?	?	1
Sheep	>10 (1)	?	6	>100 (3)	2	2	10 (4)	3	1
Camelid	VH 42 (1)								
	VH 50 (1)	10	6	?	?	2	?	?	?
Bat	>250 (5)	?	13	?	?	?	?	?	?
Opposum	12	?	?	30 (3)	6	6	35 (4)	>2	1
Platypus	25 (1)	>5	7	15-25 (2)	6	4	?	?	?

\* Number of families (F) of variable region genes.

\*\* J<sub>L</sub>-C<sub>L</sub> occurs as duplicons (see Fig. 2A).

Table 1. Variable region gene duplication among mammalian antibody genes

Species	IgM (C $\mu$ )	IgD (C $\delta$ )	IgG (C $\gamma$ )	IgE (C $\epsilon$ )	IgA (C $\alpha$ )	C <sub>L</sub>	C <sub>K</sub>
Human	1	1	>4 1*	>1 1*	2	>4 3*	1
Mouse	1	1	4	1	1	>3 1*	1
Rat	1	1	4	1	1	1	?
Rabbit	1	0	1	1	13	8	2
Swine	1	1	6	1	1	>3	1
Horse	1	1	7	1	1	4	1
Cattle	1	1	3	1	1	4	1
Sheep	1	1	>2	1	1	>1	1
Camel	1	?	>3 4*	?	?	2	1
Cat	1	?	>2	1	?	>1	1
Dog	1	1	4	1	1	>1	1
Bat	1	1**	1 5**	1	1	?	?
Opposum	1	0	1	1	1	6	1
Platypus	1	0	2	1	2	4	?

\* Additional pseudogenes

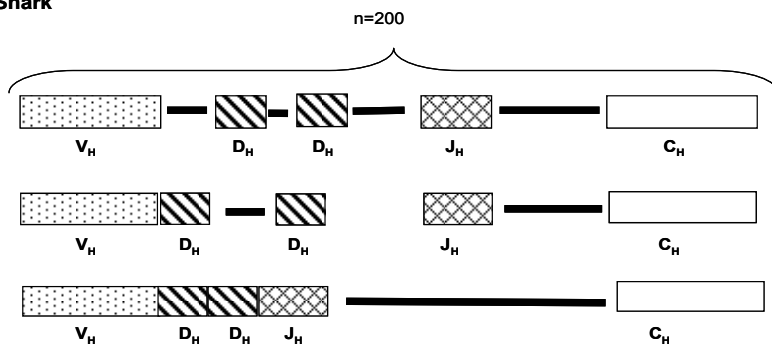
\*\* Varies with species

Table 2. Constant region gene duplication among mammalian antibody genes

## 2.2 The Ig loci in sharks and the chicken are differently organized

Figure 3 illustrates examples of the organization of V-D-J-C segments in three different shark species which are organized as repeating cassettes rather than in translocon fashion. Interestingly, Figure 2A also shows that an apparent evolutionary remnant of this form of organization is still found in the lambda light chain locus of mammals. In the shark, an entire cassette is used for encoding an antibody; recombination among cassettes is unusual. Furthermore, segments within the cassettes of certain sharks are fused in the genome so recombination (Section 4) does not occur. It is believed that the tandem repeat system of sharks later evolved into the translocon system (Marchalonis et al., 1998). In the translocon system, recombination among the various V, D, and J segments can occur and the rearranged VDJ is later spliced to a C region exon (Fig. 4A).

### Shark



### Chicken

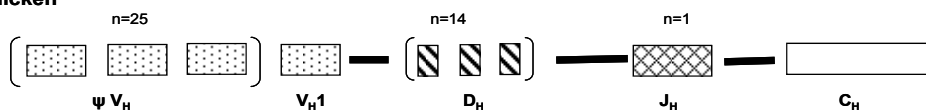


Fig. 3. Organization of heavy chain loci in sharks and chicken. Three different types of clusters are shown for sharks, some in which VDJ s are fused in the genome; n= number of repeating clusters. Modified from Dooley & Flajnik, 2006. In the diagram for chicken, the number (n) of gene segments of each type is indicated. Only V<sub>H</sub>1 of chicken is a functional V<sub>H</sub> gene.

The chicken also displays a translocon system but there is only one functional V<sub>H</sub> (and one Vλ; not shown), multiple highly similar D<sub>H</sub> segments and only one J<sub>H</sub>. All V<sub>H</sub> genes upstream of V<sub>H</sub>1 in the chicken are pseudogenes (Fig. 3; Ratcliffe, 2006). These pseudo V<sub>H</sub> genes are used in SGC to create the chicken antibody repertoire (Reynaud et al., 1987; Ratcliffe, 2006).

## 3. Duplication and diversification of Ig genes

### 3.1 V<sub>H</sub> genes display evidence of duplication and genomic gene conversion

Genomic gene conversion was originally described in yeast (Meselson & Radding, 1975; Szostak et al., 1983) and is a form of non-homologous recombination in which the end result

is that a segment of one gene is “translocated” to another gene. When this process is combined with gene duplication, an array of modified duplicons results. Figure 4 shows the VH gene sequences for swine and Figure 5 the VH3 genes of the little brown bat (*Myotis lucifugus*). We have used color-coding to show that in swine, there are only four different FR1 sequences among the 24 known VH genes, only two FR2 sequences but a larger number of different FR3 sequences, although five are often shared (Fig. 4). Also shown is that CDR regions are often shared. Assuming these genes are the result of a combination of duplication and genomic gene conversion, VHT could be derived from VHE with CDR2 and FR3 translocated from VHF. A similar pattern of shared gene segments is seen among the VH3 germline gene repertoire of the little brown bat (Fig. 5). As shown, many of these share common FR1 sequences, a smaller number share CDR1, FR2 and CDR2 while the greatest diversity is seen in FR3 (Bratsch et al., 2011). This pattern of similarity among duplicated Ig genes is also seen in human and mouse, suggesting that after duplication and genomic gene conversion, the 3' segment was subjected to a higher rate of germline mutation and selection. We believe these examples support the hypothesis that the polygeny of VH took place by a combination of gene duplication and genomic gene conversion. It is well-documented that within a sublocus, intralocus duplication of segments containing several genes also occurs. This is illustrated for the C-region sublocus for humans and rabbits (Fig. 2B). The same phenomenon occurs in the VH sublocus of mice (Retter et al., 2007; Johnston et al., 2006) humans (Matsuda et al., 1990) and in swine (Eguchi-Ogawa et al., 2010). For example, the genomic segment in swine that contains VHA, VHB and VHE has been duplicated to yield VHA\*, VHB\* and VHF.

Genebank #	V <sub>H</sub> gene	FR1	CDR1	FR2	CDR2	FR3
AF064686	V <sub>H</sub> A	EEKLVESGGGLVQPGGSLRLSCVGS	STYIN	WVRQAPGKLEWLA	AISTSGG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AB513624	V <sub>H</sub> A*	EEKLVESGGGLVQPGGSLRLSCVGS	STYIN	WVRQAPGKLEWLA	AISTSGG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AF064687	V <sub>H</sub> B	EEKLVESGGGLVQPGGSLRLSCVGS	DNAFS	WVRQAPGKLEWVA	AIASSDYDG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AB513624	V <sub>H</sub> B*	EEKLVESGGGLVQPGGSLRLSCVGS	DYAFS	WVRQAPGKLEWVA	AIASSDYDG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AF064688	V <sub>H</sub> C	EEKLVESGGGLVQPGGSLRLSCVGS	SYEIS	WVRQAPGKLEWLA	GIYSSGG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
	V <sub>H</sub> D	EEKLVESGGGLVQPGGSLRLSCVGS	SYEIS	WVRQAPGKLEWVA	DICSGG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AF064689	V <sub>H</sub> E	EEKLVESGGGLVQPGGSLRLSCVGS	SYAVS	WVRQAPGKLEWLA	GIDSGSYG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AF064690	V <sub>H</sub> F	EEKLVESGGGLVQPGGSLRLSCVGS	SYGVG	WVRQAPGKLESLA	SIGSGSYIG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
DQ886395	V <sub>H</sub> G	EEKLVESGGGLVQPGGSLRLSCVGS	SYGMS	WVRQAPGKLEWLA	GIDSGSYG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
DQ886392	V <sub>H</sub> H	EEKLVESGGGLVQPGGSLRLSCVGS	SYPIG	WVRQAPGKLEWLA	SIGGRYRG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AY911501	V <sub>H</sub> J	EEKLVESGGGLVQPGGSLRLSCVGS	SYAVE	WVRQAPGKLEWLA	SIGSGSYIG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AF064692	V <sub>H</sub> K	EEKLVESGGGLVQPGGSLRLSCVGS	SSPIG	WVRQAPGKLEWLA	SIGSGSYG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AY911500	V <sub>H</sub> L	EVKLVESGGGLVQPGGSLRLSCVGS	SYAVS	WVRQAPGKLEWLA	AIYSSGG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AF321841	V <sub>H</sub> N	EEKLVESGGGLVQPGGSLRLSCVGS	SYMS	WVRQAPGKLEWLA	GIYSSGG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AF321842	V <sub>H</sub> O	EEKLVESGGGLVQPGGSLRLSCVGS	SYPIG	WVRQAPGKLEWLA	AISTSGG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AF321844	V <sub>H</sub> P	EEKLVESGGGLVQPGGSLRLSCVGS	SYEIS	WVRQAPGKLEWLA	AISTSGA	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AF321845	V <sub>H</sub> R	EVKLVESGGGLVQPGGSLRLSCVGS	SYPIG	WVRQAPGKLEWLA	CIYSSGG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AF321846	V <sub>H</sub> S	EEKLVESGGGLVQPGGSLRLSCVGS	SYNMI	WVRQAPGKLEWLA	YITSSGG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AF321847	V <sub>H</sub> T	EEKLVESGGGLVQPGGSLRLSCVGS	SYAVS	WVRQAPGKLESLA	SIGSGSYIG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AF321848	V <sub>H</sub> U	EEKLVESGGGLVQPGGSLRLSCVGS	SYEIS	WVRQAPGKLEWLA	AIGCGSYG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AF321849	V <sub>H</sub> V	EEKLVESGGGLVQPGGSLRLSCVGS	STYIN	WVRQAPGKLEWVA	AIASSDYDG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AY911502	V <sub>H</sub> X	EEKLVESGGGLVQPGGSLRLSCVGS	SYGVG	WVRQAPGKLEWLA	GIYSSGG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
AY911504	V <sub>H</sub> ZZ	EEKLVESGGGLVQPGGSLRLSCVGS	SYMS	WVRQAPGKLEWLA	CIYSSGG	STYYADSVKGRFTISRDNQNTAYLQMNSLR
DQ886393	V <sub>H</sub> Y	EEKLVESGGGLVQPGGSLRLSCVGS	SYEIS	WVRQAPGKLEWLA	AISTSGG	STYYADSVKGRFTISRDNQNTAYLQMNSLR

Fig. 4. Deduced amino acid sequences for the framework (FR) and combinatorial determining regions (CDR) of germline porcine VH genes. Regions of FR and CDR regions that are shared among genes are color-coded. Those sequences that are not colored indicate segments with sequences that differ by one or a few changes that are not shared by other sequences.

Genebank	# group	FR1	CDR1	FR2	CDR2	FR3
GQ923685	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	DYSMN	WVRQAPGKLEWVA	YTSYSGSNPI	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923622	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	EYGMN	WVRQAPGKLEWVS	YISFSGSGNI	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923683	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	EYGMN	WVRQAPGKLEWVS	YISGDSSTDI	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923616	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	EYGMN	WVRQAPGKLEWVA	YISFSGSGNI	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923644		LVESGGGLVQPGGSLRLSCAASGFTFS	NYDMH	WVRQAPGKLEWVA	HMTDGSQK	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923675	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	NYGMN	WVRQAPGKLEWTA	YTSDDGNPI	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923628		LVESGGGLVQPGGSLRLSCAASGFTFS	NYVMN	WVRQAPGKLEWVA	SISDGSYYI	YYGEAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923647	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	NYVMN	WVRQAPGKLEWVA	YISDGSYYI	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923618	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SNSMN	WVRQAPGKLEWVA	LISDGGGST	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923650		LVESGGGLVQPGGSLRLSCAASGFTFS	SNSMN	WVRQAPGKLEWVG	IISTDGGTT	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923681	3-1	LVESGGGLVQPGGSLRLSCAASGFTFS	SSMMV	WVRQAPGKLEWVS	LINPDGSGT	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923662	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYDMN	WVRQAPGKLEWVA	LISTDGGST	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923669	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYDMN	WVRQAPGKLEWVS	LISFSGGSA	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923625	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYDMN	WVRQAPGKLEWVS	LISFSGGSA	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923612	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYDMN	WVRQAPGKLEWVA	YISSASNTI	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923621	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYDMS	WVRQAPGKLEWVS	ALISNGGST	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923679		LVESGGGLVQPGGSLRLSCAASGFTFS	SYGMH	WVRQAPGKLEWVA	YQYISSDGRNYI	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923617		LVESGGGLVQPGGSLRLSCAASGFTFS	SYGMH	WVRQAPGKLEWVA	YQYISSDGRNYI	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923658	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	SYGMH	WVRQAPGKLEWVS	RIGSDGRSYI	HYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923672	3-4	LVESGGGLVQPGGSLRLSCAASGFTFS	SYGMN	WVRQAPGKLEWVS	GVSSIGGTT	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923613	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	SYGMN	WVRQAPGKLEWVS	LISDGGSTI	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923657	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYHNM	WVRQAPGKLEWVA	FISNGGGST	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923638		LVESGGGLVQPGGSLRLSCAASGFTFS	SYQMH	WVRQAPGKLEWVE	LISDGGSTI	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923629	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYSDM	WVRQAPGKLEWVA	YISSASSTI	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923642	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYSMN	WVRQAPGKLEWVA	VISDGGSTI	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923684	3-1	LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMD	WVRQAPGKLEWLC	RMPDGGST	HYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923619	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMH	WVRQAPGKLEWVS	RISDGGSTI	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923632	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMH	WVRQAPGKLEWVS	LISDGGSTI	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923668	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMH	WVRQAPGKLEWVS	LISDGGSTI	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923640		LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMH	WVRQAPGKLEWVS	ALISDGGST	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923627		LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMS	WVRQAPGKLEWVA	HISDGGST	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923663	3-1	LVESGGGLVQPGGSLRLSCAASGFTFS	SYWNY	WVRQAPGKLEWLC	RMPDGGST	HYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923623	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	SYWNY	WVRQAPGKLEWVA	SISDGSYYI	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923680		LVESGGGLVQPGGSLRLSCAASGFTFS	DYVNH	WVRQAPGKLEWVS	DFRCDGGST	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923635		LVESGGGLVQPGGSLRLSCAASGFTFS	GYWIS	WVRQAPGKLEWVS	DISDGSSTI	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923636		LVESGGGLVQPGGSLRLSCAASGFTFS	NYDMH	WVRQAPGKLEWVA	HMTDGSQK	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923673	3-1	LVESGGGLVQPGGSLRLSCAASGFTFS	SNNMH	WVRQAPGKLEWLC	RMPDGGST	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923655	3-1	LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMH	WVRQAPGKLEWLC	QINSDDNTI	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923615	3-1	LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMH	WVRQAPGKLEWLC	QINSDDNTI	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923676	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMS	WVRQAPGKLEWVS	HISDGGST	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923631		LVESGGGLVQPGGSLRLSCAASGFTFS	SYGMS	WVRQAPGKLEWVS	GVSSIGGTT	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923611	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMS	WVRQAPGKLEWVA	LISDGGST	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923667	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	YGMN	WVRQAPGKLEWVS	GISTDGGST	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923654	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMH	WVRQAPGKLEWVS	LISDGGST	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923645	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYSDM	WVRQAPGKLEWVA	YISSASNTI	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923660	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYSDM	WVRQAPGKLEWVA	YISSASNTI	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923677	3-6	LVESGGGLVQPGGSLRLSCAASGFTFS	NYSDM	WVRQAPGKLEWVT	RVSFKPGMTQ	WYAPAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923648	3-6	LVESGGGLVQPGGSLRLSCAASGFTFS	NYSDM	WVRQAPGKLEWVT	RVSFKPGMTQ	WYAPAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923671	3-6	LVESGGGLVQPGGSLRLSCAASGFTFS	NYSDM	WVRQAPGKLEWVA	RVSFKPGMTQ	WYAPAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923626		LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMH	WVRQAPGKLEWVS	LVPAGGST	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923653		LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMH	WVRQAPGKLEWVS	LINPAGGST	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923674	3-1	LVESGGGLVQPGGSLRLSCAASGFTFS	SSMD	WVRQAPGKLEWLC	RINPDGGST	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923652		LVESGGGLVQPGGSLRLSCAASGFTFS	SYTNE	WVRQAPGKLEWVA	VISYNGNT	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923661	3-5	LVESGGGLVQPGGSLRLSCAASGFTFS	SYTNE	WVRQAPGKLEWLC	ALTANGDS	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923624	3-5	LVESGGGLVQPGGSLRLSCAASGFTFS	SYTNE	WVRQAPGKLEWLC	ALTANGDS	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923633	3-5	LVESGGGLVQPGGSLRLSCAASGFTFS	SYTNE	WVRQAPGKLEWLC	ALTANGDS	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923637	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYSDM	WVRQAPGKLEWVA	YISSASSTI	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923614	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	EYGMN	WVRQAPGKLEWVS	YISGDSSTDI	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923651		LVESGGGLVQPGGSLRLSCAASGFTFS	DYEMN	WVRQAPGKLEWVS	RITSGGST	HYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923659	3-1	LVESGGGLVQPGGSLRLSCAASGFTFS	SYNNT	WVRQAPGKLEWLC	EINPDGGST	HYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923646		LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMS	WVRQAPGKLEWVS	FVTDGGST	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923630		LVESGGGLVQPGGSLRLSCAASGFTFS	GYWIS	WVRQAPGKLEWVS	DISDGSSTI	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923682		LVESGGGLVQPGGSLRLSCAASGFTFS	GYWIS	WVRQAPGKLEWVS	DISDGSSTI	YYAASVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923649	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	SYPMN	WVRQAPGKLEWVA	LISDGGST	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923634		LVESGGGLVQPGGSLRLSCAASGFTFS	DYVMH	WVRQAPGKLEWVT	SISDGSYYI	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923641	3-2	LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMH	WVRQAPGKLEWVS	RIGSDGSYYI	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923665	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYSDM	WVRQAPGKLEWVA	YISSASSTI	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923670	3-3	LVESGGGLVQPGGSLRLSCAASGFTFS	SYSDM	WVRQAPGKLEWVA	YISSASNTI	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923639	3-4	LVESGGGLVQPGGSLRLSCAASGFTFS	SYGMN	WVRQAPGKLEWVS	GVSSIGGTT	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923656	3-4	LVESGGGLVQPGGSLRLSCAASGFTFS	SYGMN	WVRQAPGKLEWVS	GVSSIGGTT	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923620		LVESGGGLVQPGGSLRLSCAASGFTFS	DYVMH	WVRQAPGKLEWVT	LIRNKANGHT	HYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923664		LVESGGGLVQPGGSLRLSCAASGFTFS	SYSDM	WVRQAPGKLEWVS	TIHTDGGST	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923643	3-6	LVESGGGLVQPGGSLRLSCAASGFTFS	NYSMY	WVRQAPGKLEWVA	GVSKPTGKQ	WYAPAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923678		LVESGGGLVQPGGSLRLSCAASGFTFS	SYWMS	WVRQAPGKLEWVS	LITNGGST	YYANAVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR
GQ923666	3-5	LVESGGGLVQPGGSLRLSCAASGFTFS	SYTNE	WVRQAPGKLEWLC	ALTANGDS	YYADSVKGRFTISRDNKNTLYLQMSSLRADETAVYYCAR

Fig. 5. The deduced amino acid sequences for the framework (FR) and combinatorial determining regions (CDR) for 75 members of the VH3 family from the little brown bat (*M. lucifugus*). Shared sequences are color-coded as in Fig. 4. From Bratsch et al., 2011.

### 3.2 Duplication and deletion of Ig genes in the C-region sublocus

Figure 2B illustrates that in humans, a major segment of the C-region sublocus has been duplicated resulting in two IgEs, two IgAs and two pair of IgG genes (Lefranc et al., 1982; Flanagan & Rabbitts, 1982). A  $C\gamma$  pseudogenes separates them. In mammals, the target of duplications in the C-region has been those genes encoding IgG (or IgA in rabbits; Fig. 2B; Table 2). The duplication process in the CH1 region suggests it also occurred together with genomic gene conversions to produce an array of modified  $C\gamma$  genes (Fig. 6). In the example provided the IgG1 and IgG2 alleles share a common CH1 domain that is also found in the IgG4 alleles. The allelic variants of IgG1 and IgG4 differ only in their hinge exons. However, IgG1<sup>a</sup> and IgG4<sup>a</sup> have the same hinge as do IgG1<sup>b</sup> and IgG4<sup>b</sup>. The difference between IgG1 and IgG4 is in the CH3 domain which is not shared with any other  $C\gamma$  subclass gene. Another example is IgG5<sup>a</sup> and IgG6 which share a common CH1 and hinge exons. IgG5<sup>a</sup> also has the same CH2 exon as IgG6<sup>b</sup>, but the difference is in CH3. Thus Fig. 6 also shows that apparent genomic gene conversion events also involve allelic variants as well as entire genes. The pattern shows that the CH1, hinge and CH2 of IgG1<sup>a</sup> and IgG4<sup>a</sup> were derived from the same ancestral  $C\gamma$  gene but the IgG1<sup>b</sup> and IgG4<sup>b</sup> were derived from another ancestral gene. The reverse effects of genomic gene conversion may explain certain heterozygous  $C\gamma$  deletions (Migone et al., 1984; Keyeux et al., 1989). Some swine lack certain  $C\gamma$  genes (Butler et al., 2009a) and deletions in the human  $C\gamma$  sublocus are well documented (LeFranc et al., 1983a; Rabbani et al., 1995).

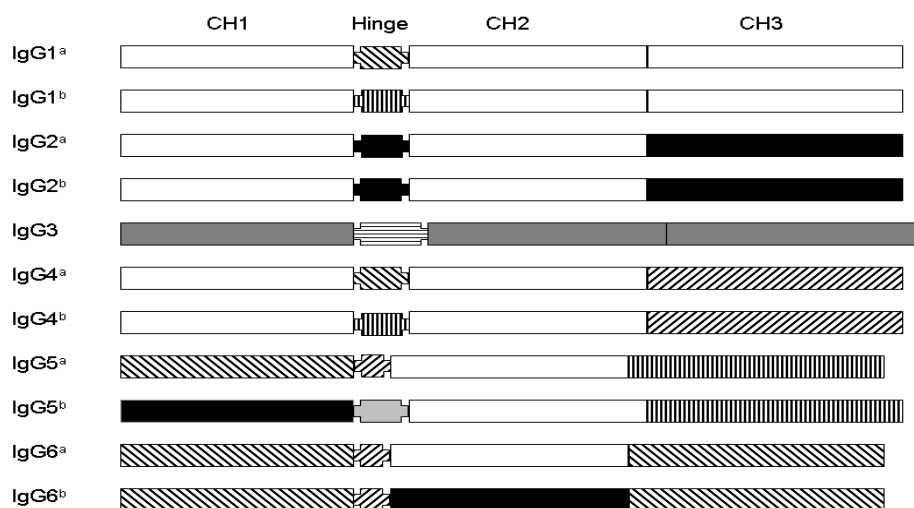


Fig. 6. Alignment of the constant region domains of the porcine  $C\gamma$  genes. Regions of >95% homology are designated with the same texture. Superscripts in the gene designation denote allotypic variants. From Butler et al., 2009a.

As we have shown elsewhere, porcine IgG3 has a gene structure which is most similar to the consensus  $C\gamma$  genes of other mammals and therefore is closest to the ancestral  $C\gamma$  gene of all mammals (Butler et al., 2009a). IgG3 in humans, mice and swine all occupy the same 5' position which is immediately downstream of  $C\delta$  (Eguchi-Ogawa et al., 2010). Our studies indicate that the remainder of the porcine  $C\gamma$  genes were derived from an ancestral  $C\gamma$  that



diverged early from IgG3 (Butler et al., 2009a). We hypothesize that the duplication/diversification of  $C\gamma$  subclass genes in mice and humans followed the same pattern.

## 4. Mammalian antibody repertoires result from somatic events

### 4.1 Somatic gene segment recombination characterizes B cell lymphogenesis

B lymphocytes, named because they form in the Bone marrow or the chicken Bursa of Fabricius, are the cells that synthesize and secrete antibodies. This developmental process occurs in what are called “primary lymphoid tissues”. These include the bone marrow, the chicken bursa, fetal liver, yolk sac and according to some, certain hindgut lymphoid tissues of artiodactyls. Among lower vertebrates, other tissues like the “head kidney” (pronephros), epigonal organ and Leydig organ are involved in this process (Solem & Stenvik, 2006; Rumpf et al., 2002; Dooley & Flajnik, 2006).

Somatic recombination is illustrated in Figure 7A. This process is mediated by Recombinase Activation Genes (RAGs) as well as a variety of DNA repair and ligation enzymes. In the heavy chain locus this process first involves recombination of one J region gene segment and one D region gene segment. This event also produces a circular DNA product containing the intervening DNA sequence that is excised and is known as a signal joint circle (Fig. 7B). Single joint circles are diagnostic evidence that B cell lymphogenesis has recently occurred since this nuclear product is rapidly degraded. The rearrangement process then proceeds to the rearrangement of the DJ unit with some V gene segment and generation of another signal joint circle (Fig. 7A). The selection of the J, D, and V gene segments is poorly understood and will be discussed in Section 5. A similar series of events occurs among segments in the light chain loci except that there are no D segments involved.

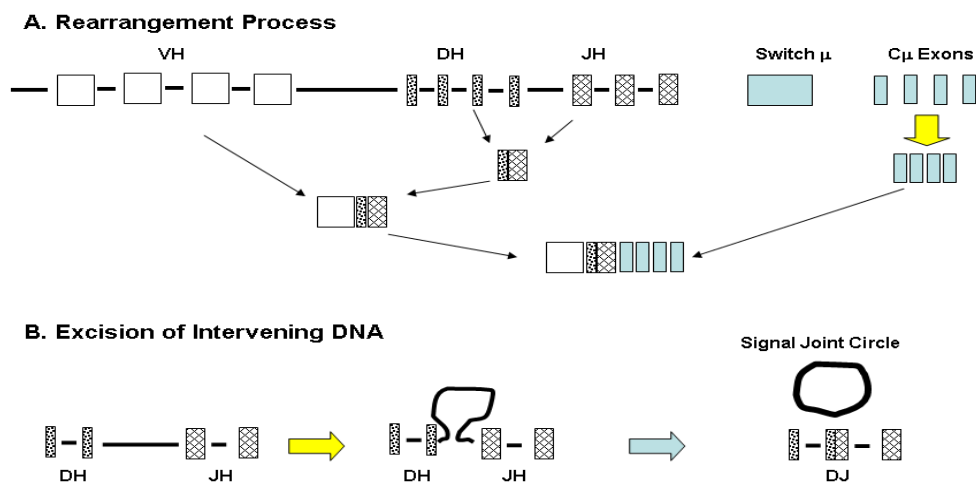


Fig. 7. The somatic rearrangement process among the gene segments of the variable heavy chain locus of mammals. A. Sequential rearrangement of D to J, then DJ to V and finally splicing of the primary transcript for VDJ to the exons encoding the C-region of IgM. B. Generation of a signal joint circle during the excision of intervening DNA during recombination of D and J.

The rearranged VDJ (heavy chain) and VJ (light chains) rearrangements are then transcribed and the primary transcript spliced to some set of C-region exons in the heavy and light chain loci respectively. For the heavy chain this is initially IgM in all higher vertebrates (Fig. 7A). The resulting VDJ-C and VJ-C transcripts are then translated into the light and heavy polypeptide chains that combine to form the complete antibody molecule (Fig. 1B). As shown in Fig. 1A and discussed above, the antigen-binding site is located in the peptide loops from the VH and VL domains that coalesce at the end of these V-domain and that contain the CDRs; three from the VDJ and three from the light chain VJ rearrangements. CDR1 and CDR2 of the VH and VL are encoded within the germline genes whereas the CDR3 region is the result of the combining of V-D-J segments (heavy chain) and V-J (light chains; Fig. 8). The CDR3 region of the heavy chain, hereafter designated HCDR3, is considered most important to the specificity of the binding site (Amit et al., 1986; Padlan, 1996; Xu & Davis 2000; Mageed et al., 2001). In fact the same set of V-genes encoding CDR1 and CDR2 can theoretically and actually contribute the binding site for antibodies of quite different specificity in the context of different HCDR3 regions (Thomson et al., 2011; Ichiyoshi & Casali, 1994) as might be envisioned from a comparison of the CDR3 sequences shown in Fig. 8.

FR3	5' N-add	D <sub>H</sub> A	3' N-add	J <sub>H</sub>	FR4
TGNGCNA		<u>GAATGCTATAGCTATGGTGTCTAGTTGCTATATGATGAC</u>		ATTACTATGTTTGGATCTCTGGGGGCCA	
TGTGCAAGT		<u>TGCTATAGCTATGGTGTCTAGTTGC</u>	TTTTGGACAAGATCA	TACTATGCTATGGATCTCTGGGGGCCA	
TGTGCAAGA	GGCTGTTTTC	<u>GCTATAGCTATGGTGTCTAGTTGCTAT</u>	GATGTCG	ACTATGCTATGGATCTCTGGGGGCCA	
TGTGCAA	CAGGCGAT	<u>TGCTATAGCTA</u>	GGTGTCTAGTTGCACCGGGATG	GCTATGGATCTCTGGGGGCCA	
TGTGCAA	TT	<u>GCTATAGCTATGGTGTCTAGTTG</u>	T	TATGGATCTCTGGGGGCCA	
TGTGCAA	CAGAG	<u>TGCTATAGCTATGGTGTCTAGTTGCTATAT</u>	GTATGC	TATGGATCTCTGGGGGCCA	
TGTGCAAGA	G	<u>ATAGCTATGGTGTCTAGTT</u>	ACCCCTC	TATGGATCTCTGGGGGCCA	
TGTGC	CCAG	<u>GCTATAGCTATGGTGTCTAGT</u>	CCAGGATG	TGGATCTCTGGGGGCCA	
TGTGCAA	CAGGC	<u>ATAGCTATGGTGTCTAGTTGCTAT</u>	GAAGA	TGGATCTCTGGGGGCCA	
TGTGCAAG	GTCC	<u>AATGCTATAGC</u>	TCCGTTGGTGTAGTGTCTATGGTTACCCCTGGGGTTATGTTGCTG	TGGATCTCTGGGGGCCA	
TGTGCAA	TT	<u>GCTATAGCTATGGTGTCTAGTT</u>	AGATC	GGATCTCTGGGGGCCA	

Fig. 8. The diversity of HCDR3 sequences resulting from the recombination of the same VH, DH and JH segments. Remnants of the DH germline segments are underlined. The 5' and 3' nucleotide additions are indicated. TGG is the codon for the invariant tryptophan found in the JH gene segments of all mammals while TGT is the codon for C that is nearly invariant in the FR3 of all VH3 family genes.

## 4.2 Maturation of the antibody repertoire involves class switch and somatic hypermutation

All immunologists, immunopathologist and physicians in specialties such as rheumatology know that most Igs are IgG (serum) or IgA (secretions). This means that the rearrangements involved in B cell lymphogenesis that initially favors the expression of IgM (Fig. 7A) switch to these isotypes. After environmental exposure, the concentration of the major Igs in serum is elevated 100-300 fold compared to newborn piglets or those reared in germfree isolators (Fig. 9A). The transition from newborn to conventionally-reared young adults favors IgG in serum (Fig. 9A) and IgA in secretions (Butler et al., 2011a). This change involves class switch recombination (CSR) which is mediated by activation-induced cytidine deaminase (AID) of the APOBEC family which facilitates the splicing of RNA encoding the rearranged VDJ to transcripts encoding IgG and IgA rather than IgM. This maturation process typically occurs in tandem with somatic hypermutation (SHM) of the rearranged VDJ or VJ prior to their transcription. SHM is another mechanism for repertoire diversification and is triggered

when the neonate encounters environmental antigen (Fig. 9B). Since both CSR and SHM occur simultaneously, it is not surprising that both are mediated by AID and that AID is also correlated with SGC (Withers et al., 2005; Arakawa et al., 1996). These events occur in germinal centers (GCs) of secondary lymphoid tissues after exposure to environmental antigen. GCs are found only in mammals and birds (Yasuda et al., 2003; Vigliano et al., 2006; Du Pasquier et al., 2000). Although lacking GCs, there is SHM and CSR in *Xenopus* (Marr et al., 2007) although it may be less efficient. However, for these events to occur, the naïve immune system must first or simultaneously be exposed to Pathogen Associated Molecular Patterns (PAMPs) that are recognized by a variety of innate immune system receptors. This dependence was demonstrated using the isolator piglet model (Butler et al., 2002; 2005; 2009b; Butler & Sinkora 2007).

SHM is not random across the entire rearranged VDJ-C transcript. Rather it is largely concentrated in the CDR regions of the rearranged VDJ or VJ segments (Fig. 9C). This is generally believed to result from selection of B centrocytes in GCs rather than specific targeting. Although Fig. 9C only shows the accumulation of somatic mutations in CDR1 and CDR2, the same occurs in CDR3. As discussed previously, the CDRs are those segments of the encoded protein that coalesce to form the antibody binding site (Fig. 1A; Fig 8). There is little evidence to suggest that SHM proceeds downstream from segments of transcript that begins with the codon for the invariant tryptophan in FR4 (Fig. 8) or to sequences further downstream in the C-sublocus.

#### **4.3 The association of VH genes and VH- VL pairing in generation of specific antibodies**

Much of the early studies on antibody specificity that appeared when VH or VL polygeny became known, attempted to correlate particular response to the use of certain VH or VL genes. We do not review that literature here but do provide a few examples. Cerato et al., (1997) studied hybridomas to show a lack of correlation between VH usage and specificity while Mo and Holmdahl (1996) show that mAbs to different epitopes used the same VH/Vk combinations. Boffey et al., (2004) showed that only 6/15 anti-LPS mAbs used the same VH gene (VH7183.3b). These observations should not be surprising considering the importance of HCDR3 in the specificity of antibodies (see Section 4.3; 6.2). Lavoie et al., (1997) showed that nearly all mAbs to HEL use VH36-60 but differ in affinity because of SHM or HCDR3 differences. The antibody binding site involves CDRs (including CDR3) of both H and L chains (Fig 1B); this has been shown by separation and reassociation experiments. These experiments show that binding site specificity depends on both H and L chains even for antibodies specific for the same hapten since heterologous light chains seldom restore the full binding site (Kranz & Voss, 1981). This mutual dependence is also demonstrated by the non-random pairing found in antibodies of certain specificity such as to the capsular polysaccharides of *S. pneumoniae* (Thomson et al., 2011). Further evidence for the effect of H-L pairing comes from studies of autoantibodies in a phenomenon called "receptor editing" (deWildt et al., 1999). This *in vivo* phenomenon involves reactivation of recombinase activity in lymph nodes resulting in the replacement of the light chain with a new one. In this way, B cells expressing autoreactive BCRs acquire a new light chain which alters their specificity and removes or diminishes their autoreactivity apoptotic elimination (Tiegs et al., 1993; Gay et al., 1993).

While L-H pairing is important for binding site specificity, there are situations in which light chains are not needed to form an antibody binding site. The best known examples are the

naturally occurring single chain antibodies of the camelid group and some sharks (Hamers-Casterman et al., 1993; De Genst et al., 2006; Dooley et al., 2003; Diaz et al., 2002; Nguyen et al., 2002). Based on the convenience of producing single chain antibodies from these species for therapy and the evidence that the HCDR3 domain plays the major role in forming the antibody binding site (see Section 6.2) there have also been various attempts to develop synthetic single chain antibodies or “camelized” antibodies (Janssens et al., 2006; Reiter et al., 1999; Davies & Riechmann, 1995). Among the camelids, single chain antibodies use a separate set of VH genes (called VHH) that encode a much larger portion of the binding sites than the conventional VH genes which compensates for the lack of a light chain. This topic has been recently reviewed (Muyldermans et al., 2009). We mention these single chain antibodies here because we believe they further support the role played by HCDR3 in binding antigen and diminishes the value of polygeny of conventional VH genes (Section 6.2). It also shows that the extensive and universal V $\kappa$  and V $\lambda$  polygeny among mammals (Table 1) is unnecessary.

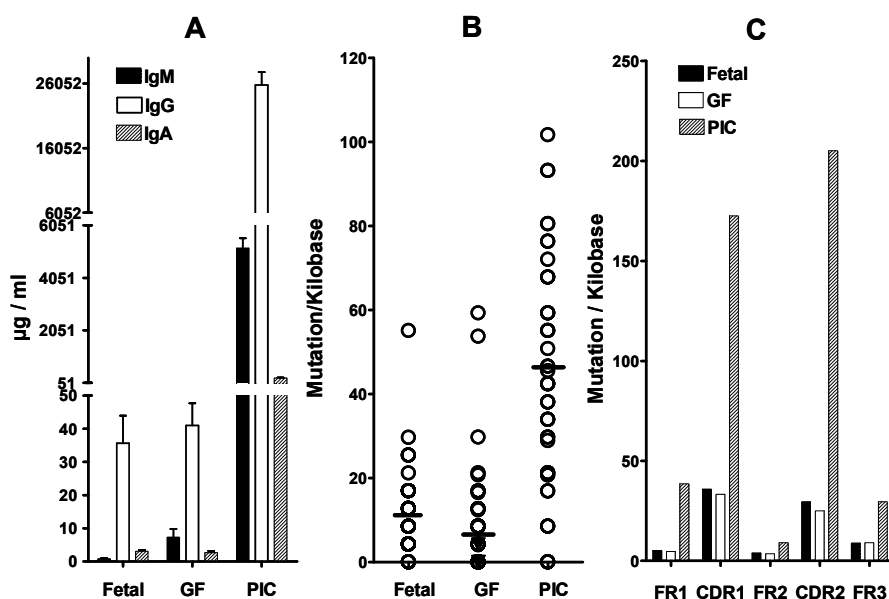


Fig. 9. The effect of antigen exposure on: A. Serum Ig levels; B. Frequency of SHM and C. Accumulation of somatic mutations in various segments of the VH genes. Germfree piglets are reared in isolators for 5 weeks and their only contact with potentially foreign antigen is food protein. PIC=conventionally-reared young pigs that are heavily antigenized through colonization and also infected with nematodes. The horizontal line (9B) in the scattergram is the mean frequency of SHM. SHM is significantly greater in PIC piglets than in fetal and germfree piglets. In 9C SHM accumulates in CDR regions as opposed to FR regions that encode the  $\beta$ -pleated “staves” of the  $\beta$ -barrel (Fig. 1A).

## 5. Patterns of V, D and J gene segment usage

### 5.1 VH usage is biased to favor certain VH genes

Table 1 shows that higher vertebrates have many duplicated V-region gene segments available for use in the formation of their antibody repertoire using the recombinatorial process illustrated in Figure 7. Humans have available ~ 100 VH segments, ~30 DH segments and 9 JH segments (Fig. 2A). By contrast, swine have <30 VH genes belonging to a single family (Fig. 4), only two functional DH segments and like the chicken (Fig. 3) one functional JH segment (Sun et al., 1994; Butler et al., 1996; Eguchi-Ogawa et al., 2010). While the ancestral VH3 family (Schroeder et al., 1990) dominates the V-region loci of many species, the ~100 VH genes of mice and human belong to 14 and 7 different families respectively (Table 1).

Usage of VH genes in rabbit is biased to the most 3' VH gene, which accounts for 90% of VH usage in the pre-immune repertoire although there are >100 VH genes in the rabbit repertoire (Currier et al., 1988; Table 2). In humans there is bias for V3-23, V3-30, V3-33 and V4-34 (Glas et al., 2000). While some suggest that VH usage is random in mice (Dildrop et al., 1985) studies on J558 usage (one-half of the mouse genome) indicates that usage is unequal and rather scattered across the entire J558 genome even in the pre-immune repertoire (Gu et al., 1991) and that usage is not affected by SHM or CSR. Foster et al., (1997) showed that while most Vk genes were used, usage was non-random and the same was true for Jk. Sheehan et al., (1993) showed that fetal VH usage can differ from 0.1 to 1.0 but that most 5' VH genes are underrepresented. In swine VHA (IGHV4) and its near duplicate (IGHV10; see Figs. 4 & 10) account for one-third to one-half of the pre-immune repertoire (Butler et al., 2006; Eguchi-Ogawa et al., 2010; Butler et al., 2011b). Interestingly, the majority of these preferred genes in all these species belong to the ancestral VH3 family (Schroeder et al., 1990; Brezinschek et al., 1997).

### 5.2 Variable region gene segment usage is not position dependent

Early studies suggested that VH usage was biased during early stages of B cell lymphogenesis to favor the most JH proximal DH segments and the most 3' VH genes (Schroeder et al., 1987; Yancopoulos et al., 1984) but that this pattern became "normalized" in adults (Malynn et al., 1987). This concept gained support when it was found that young rabbits use their 3' most VH gene > 90% of the time and then further diversified their repertoire using upstream VH genes and SGC; perhaps a type of "developmental normalization" (Knight 1992; Becker & Knight 1990). However, additional studies in humans neither substantiated the positional "3' bias" (Matsuda et al., 1993) nor have our studies in swine (Eguchi-Ogawa et al., 2010; Fig. 10). The most 3' functional VH in swine (IGHV2) is almost never used while upstream VH15 (IGHV15) can account for ~13% of VH usage (Fig. 10). Thus, the "position hypothesis" to explain VH usage has not been universally fulfilled.

### 5.3 VH gene usage remains constant in fetal and young pigs

Vertical studies on VH usage in especially humans and mice are difficult because: (a) the V-D-J repertoire of these species is complex (Table 2) and could require up to 56 primer sets to recover all VDJ rearrangements in mouse and 42 sets for human (b) maternal regulatory factors transmitted *in utero* or via colostrum/milk can influence pre-natal and postnatal development (Wikler et al., 1980; Rodkey & Adler, 1983; Klobasa et al., 1981; Wang & Shlomchik 1998; Yamaguchi et al., 1983) and (c) control of environmental and maternal

effects is difficult or impossible in species with altricial offspring. Therefore we addressed this issue using a piglet model in which there is no transfer of maternal factors *in utero* and the influence of environmental factors postnatally on their precocial offspring can be controlled by the experimenter (Butler & Sinkora 2007; Butler et al., 2009b). Use of this model revealed that VH usage was constant during fetal life and that seven major genes accounted for >90% of the repertoire (Fig. 10; Butler et al. 2011b) while four can explain >80% of the repertoire. Interestingly in piglets exposed to viral infection, gut colonization or nematode parasites after birth, ~75% of the mutated genes used were the same seven (Fig. 10; extreme right). Furthermore, proportional usage of these genes was similar to what was seen in the pre-immune repertoire, albeit somatically mutated. Some modest changes were observed such as an increase in VHJ and decreases in VHA\* and VHN. In other words, swine seldom select other genes from their repertoire after exposure to environmental antigen, but continue to use the same VH genes that comprise ~93% of the pre-immune repertoire.

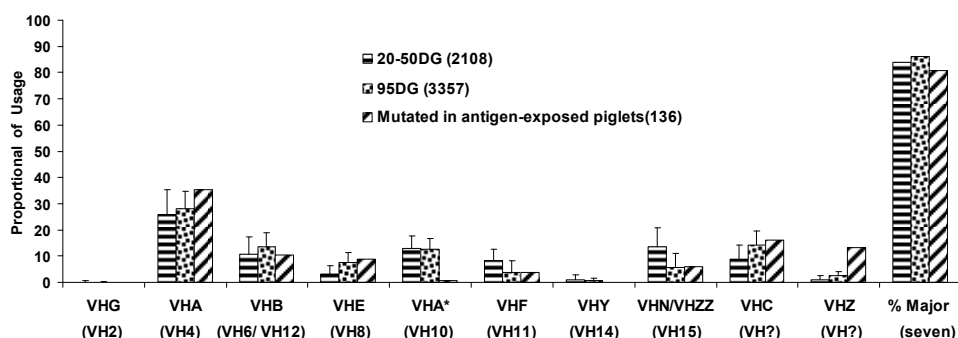


Fig. 10. VH gene usage in fetal piglets and among neonates that are antigen-exposed remains relatively constant. DG= days of gestation. The number of VH gene clones tested is given in the legend. The mutated VH genes are no longer recognized because they do not hybridize with VH gene-specific probes. Their identity must then be determined by sequencing. The bar graph on the extreme right gives the proportion of all mutated VH genes that are accounted for by the major seven genes used in the pre-immune repertoire by the fetus.

These observations should not be surprising considering that the specificity of binding site is heavily dependent on HCDR3 (Section 6.2). The HCDR3 repertoire in swine is diverse when only one VH, one DH and one JH segment are used (Fig. 8). Furthermore, the limited use of VH genes should not limit specificity, since their CDR regions can also be somatically mutated (Fig. 9B, C). As discussed in 4.1, the variants from a single VH (Fig. 8) can pair with different light chains, which can further reduce the need for large numbers of different VH genes for repertoire formation.

#### 5.4 Why is the germline VH repertoire large but usage of the repertoire limited?

Use of the piglet model demonstrates that most of the piglet VH repertoire (Fig. 4 versus Fig. 10) is seldom used to form the antibody repertoire. While corresponding vertical studied are lacking in mice and human because of the logistic and experimental reasons

discussed in the last section, the bias usage of certain VH genes reported for these species (Section 5.1) suggests the outcome might be similar if such studies could be efficiently performed. The answer to the question may reside in understanding the evolution of the genes encoding antibody specificity that generated the vast array of Ig variable region genes first seen in more primitive vertebrates and the later evolution of somatic events that would appear to have made the original polygeny unnecessary.

## **6. The value of gene duplication to adaptive immunity**

### **6.1 Adaptive versus innate immunity**

The innate immune system is the “first responder” element of immune protection for higher vertebrates and may be the sole system for most invertebrate. Innate immunity works through phagocytic and epithelial cells that bear so-called “innate immune receptors”, e.g. Toll-like (TLR) that recognize bacterial and viral entities that are not produced by eucaryotic cells of the host and are therefore considered foreign and dangerous. For vertebrates these include lipopolysaccharide, flagellin, bacterial DNA (non-methylated) and double-stranded RNA. These are referred to as PAMPs (Pathogen Associated Molecular Patterns). Products liberated from cells stimulated when their innate receptors recognize these PAMPs, then stimulate lymphocytes that leads to development of the adaptive immune system. This is demonstrated in studies using the isolator piglet model in which colonization or purified PAMPs are required for an adaptive immune response (Butler et al., 2002; 2005) and by infection with influenza virus which generates double-stranded RNA during replication (Butler, Lager, Vincent, Sun unpublished).

Unlike antibodies, the ligand binding sites of innate immune receptors do not change their specificity by any somatic process when they encounter a PAMP. The capacity for receptor modification after antigen encounter is the property of the adaptive immune system, as implied by the name. This adaptive capacity is illustrated in Figure 9 by showing there is a profound increase in Ig secretion and a shift in isotype usage (Fig. 9A) and an increase in SHM of the adaptive immune system antigen receptors after antigen exposure (Fig. 9B). As illustrated, these somatic hypermutations accumulate in those regions of the VH genes that encode the antibody binding site, i.e. the CDRs (Fig. 9C; Fig. 1A).

### **6.2 Are multiple V<sub>H</sub> genes required for host immune protection?**

Since swine use a very small number of VH genes to generate a VDJ repertoire capable of protecting the host at all ages (Sun et al., 1998; Butler et al., 2006), we did a statistical iteration to suggest that >95% of the adaptive VDJ repertoire was the result of diversity within HCDR3 (Butler et al., 2000). HCDR3 is not encoded by any particular V-region gene segment but rather by the recombination of VH-DH-JH (Fig. 7 & 8). Joining of VH-DH-JH involves exonuclease removal of nucleotides from the gene segments involved as well as nucleotide additions using deoxynucleotide transferase. The types of variations generated are illustrated in Fig. 8 which shows different HCDR3 sequences found among the recombinants that use just one VH, DH and JH segment. The importance of the diversity generated in HCDR3 was empirically shown by Xu and Davis (2000) to be sufficient to allow most adaptive immune response using a transgenic mouse given only on VH gene but an intact DH and JH genome. Studies in rabbit also show that a single VH gene is primarily used in the establishment of the antibody repertoire (Knight 1992) but following antigen encounter this VH gene can be modified by SHM as well as by SGC

(Winstead et al., 1999; Schiaffella et al., 1999). We show that in swine seven VH genes can account for 93% of the entire pre-repertoire and that the two duplicated VHA genes (which have identical CDR1 and CDR regions; Fig. 4) can alone account for 30-50% (Fig. 10). Collectively these studies raise the question as to why the VH repertoire has been so heavily duplicated while so few of these duplicons are used. The answer may be found among bats, or at least in the insectivorous microbats. *Myotis lucifugus* has >250 VH3 family genes (Fig. 5) and probably 350 total VH genes, including all families (Bratsch et al., 2011). However, SHM occurs at the frequency seen in fetal piglets (Fig. 9B). Perhaps some older mammalian orders like the Chiroptera, rely more heavily on VH polygeny than somatic modification.

### 6.3 How does duplication/ diversification in the C-region effect protective immunity

The survival value of C-region polygeny can be appreciated because antibody isotypes encoded by the exons for IgM, IgG etc have distinct biological functions (Janeway et al., 2005). Additional duplication of  $\gamma$  and  $\alpha$  genes has generated modified duplicons encoding Ig subclasses also with different important biological functions. In the case of IgG subclasses, these involve features like the ability to be recognized by different Fc receptors on phagocytic cells, transport across epithelial barriers, serum half-life differences involving FcRn and an association with antibodies of certain specificities. Duplication of  $\gamma$  genes in cattle (and other domestic ruminants) has lead to subclass IgG1 that is selectively transported by acinar epithelial cells of the mammary gland to provide essential antibodies for the survival of the newborn (Butler 1974; 1998). Neither IgG2 nor IgG3 function in this capacity. While there have been no functional studies on the many IgG subclass antibodies in horse or swine, it is possible that each of these subclass antibodies are best suited for particular activities in the same manner as described for human IgG subclasses.

The duplicated human  $\alpha$  genes also differ in a number of features. Most striking is the susceptibility of IgA1 to IgA proteases produced by many common Gram positive bacteria while IgA2 is resistance due to the lack of the 13 amino acid hinge which is the substrate for these proteases in IgA1 (Plaut et al., 1974). Differential tissue expression of rabbit IgA subclasses also suggests a division of labor (Spieker-Polet et al., 1993). IgA in swine lack the long hinge of human IgA1 and is therefore not susceptible to the classical bacterial IgA1 proteases although a protease from *H. suis* can cleave the porcine  $\alpha$ -chain (M. Mullins, K. Register, D.O. Bayles, J.E. Butler, unpub).

The “experiment of nature” is whether mammals with a deficiency in their  $\gamma$  sublocus are immunologically impaired. There are no known ruminates that lack ruminat IgG1, so there are no data; perhaps such a deficiency would be a developmental lethal. The mammals best-studied for  $\gamma$  subclass deficiencies are humans. For example, humans lacking IgG2 have a deficiency in their response to bacterial polysaccharide antigens (Hammarstrom et al., 1986). Additional deletions of  $\gamma$  genes have been described, including individuals lacking  $\gamma$ 1,  $\gamma$ 2,  $\gamma$ 4 and  $\alpha$ 1 (the 5' duplicon shown in Figure 2B). However, such individuals remain healthy and asymptomatic (Lefranc et al., 1983b; Notoaramgelo et al., 2009). Selective IgG1 deficiency which accounts for the major portion of serum IgG, is not correlated with lower serum IgG levels (Olsson et al., 1993) while some appear to be (Rabbani et al., 1995). In swine, the IgA<sup>b</sup> allotypic variant lacks a major portion of its hinge (Brown et al., 1995), yet this “defect” has not been correlated with any risk of disease; (Navarro et al., 2000).



#### 6.4 Polygeny is widespread in other loci important to immunity

There are various examples of polygeny in the immune system but in these loci SHM is not important in repertoire formation. The loci encoding the various T cell receptors are similar to those encoding the Igs. Recombination of gene segments occurs in the same manner as for Ig genes, i.e. recombination requires RAGs, DNA excision, and splicing and DNA repair enzymes (see Fig. 7). However there is no convincing evidence for SHM after segment recombination so polygeny in the TCR is theoretically more important in repertoire generation than in the Ig loci where SHM can further diversify the repertoire.

The genes encoding the major histocompatibility genes (MHC) are partially encoded by IGSF genes and determine: (a) "individualism" and self recognition as demonstrated in tissue typing and (b) recognition of peptides generated and presented by antigen-presenting cells. The MHC gene system is highly conserved among mammals with three genes encoding MHC I and 6-8 encoding MHC II. The enormous diversity of MHC is not achieved by polygeny or somatic processes but rather by an enormous degree of polymorphism of the MHC genes within the population (Janeway et al., 2005). There can be as many as 300 allelic variants of any one MHC gene.

The selective advantage of polygeny for genes encoding the MHC, innate immune receptors, the CH and C $\gamma$  subclasses and even the TCR is easy to appreciate. However, such polygeny in loci encoding the variable Ig genes in higher vertebrates is more difficult to explain if these species can generate an effective antibody response using a single VH gene (Section 6.2). This creates an enigma for VH and VL polygeny that we shall attempt to explain from an evolutionary perspective.

### 7. Conclusions: Ig polygeny and redundancy in higher vertebrates is an evolutionary vestige

#### 7.1 The case of mammalian IgD

Vestiges of genes are not unusual and mammalian IgD is an example. IgD was discovered as a myeloma protein nearly 50 years ago and its function has remained an enigma since that time (Rowe & Fahey, 1965). The considerable research funding invested to determine the function of IgD has largely generated only hypotheses. IgD is the least homologous isotype among mammals, e.g. <40%, (in part the reason it was overlooked in some mammals; Butler et al., 1996) whereas most other isotypes share 70-90% homology (Butler, 2006). IgD is even missing from the genome of some mammals (Table 2; Fig. 2B) and perhaps all birds. In mice and humans, IgD and IgM occur as dual B cell receptors but IgD-deficient mice have normal immune responses (Nitschke et al., 1993) although IgD can compensate for the loss of functional IgM (Lutz et al., 1998). While numerous studies have attempted to define a unique role for mammalian IgD, most of these have not been very convincing (Monroe et al., 1983; Liu et al., 1996; Roes et al., 1993; Vitetta et al., 1977).

Comparative immunologists have put the role of IgD into perspective beginning with the observation in catfish of a seven domain Ig with distant homology to mammalian IgD (Wilson et al., 1997; Bengten et al., 2002). This was followed by the discovery of a similar multi-domain IgD in *Xenopus* (Zhao et al., 2006) and in other teleosts (Hordvik et al., 1999; Srisapome et al., 2004; Stenvik & Jorgensen, 2000). IgD has subsequently been found in the genome of many other lower vertebrates and in protherian mammals (reviewed by Edholm et al., 2010). Collectively these studies would morph into the realization that IgD and IgM are the primordial vertebrates Ig isotypes (Ohta & Flajnik, 2006; Bengten et al., 2006).

Interesting among the studies in catfish and swine, is that IgD can be produced as a chimeric Ig using the C $\mu$ 1 domain of IgM and the various domain exons of C $\delta$  (Zhao et al., 2002; 2003). Thus, the expression of IgD in mice and humans by RNA splicing rather than classical CSR (Maki et al., 1981) has a primitive history.

Compared to the IgD of fish and *Xenopus*, the exons encoding for mammalian IgD appear to be relics. Depending on the mammal selected, the number of hinge and domain exons is highly variable (Butler et al., 2010). While switch regions are typically >3kb in length, mammalian S $\delta$  (when present) exists as a short remnant of <0.5 kb (Zhao et al., 2003). Nevertheless it appears to function in some cases in CSR in humans and swine (Zhao et al., 2002; 2003; Arpin et al., 1998; Koelsch et al., 2007; Zheng et al., 2004).

Recent findings show that basophils have abundant membrane IgD but not T cells, NK cells, dendritic cells or monocytes (Chen et al., 2009; Dawichi & Marshall, 2007) although this observation was not supported by the recent report by Karasuyami et al., (2009). Assuming the case for adventitious IgD on basophils is true, it agrees with studies in catfish showing surface IgD on granulocytes (Edholm et al., 2010). These observations would support a unique role of this ancestral Ig. In spite of the observation of IgD on human basophils, the “experiment of nature” that IgD deficiency does not impair mammalian adaptive immunity (Nitschke et al., 1993) has lead us to conclude that IgD is an evolutionary relic for mammals but persists because of its redundant value including its role as a BCR and its presence as an adventitious Ig on basophils (Lutz et al., 1998; Chen et al., 2009). Thus, the relic remains because there has been no negative selection to completely remove IgD from most mammalian genomes.

## **7.2 CH duplication/diversification provides antibodies with specialized effector function but also redundancy**

Isotype diversity in sharks and bony fish is limited to IgM and IgD (IgW in sharks). In tetrapods the CH repertoire diversified; 3 in birds and reptiles, 5-6 in amphibians and 5 in mammals (Table 2). In mammals this includes subclasses of IgG and IgA. Each major isotypes in mammals, perhaps with the exception of IgD, has some specialized effector function (Janeway et al., 2005). Perhaps the terrestrial environment offered a new challenge to survival and with the addition of homeothermia, the need for a more specialized adaptive immune system that could respond more quickly and lead to the evolution of GCs (Sections 3.2; 6.3). As discussed in Section 3-2, there is also evidence that the polygeny of C $\gamma$  genes resulted from a combination of gene duplication and genomic gene conversion. This duplication event was restricted to mammals that appeared in the “last minute before 12PM” on the evolutionary clock, appearing after mammalian speciation (Butler et al., 2009a). The subclass duplication/diversification in mammals resembles the pattern that produced V region polygeny. In sharks and bony fishes, isotype diversity is limited to IgM and IgD, while in higher vertebrates duplication/ diversification extends downstream into the CH sublocus, and in mammals, especially to the C $\gamma$  sublocus. However, IgG subclass deficiencies are not lethal defects, suggesting that even in late-evolving sites of Ig gene duplication, such duplication seem unnecessary.

## **7.3 Somatic mechanisms render much of the polygeny in V $H$ , V $\lambda$ and V $\kappa$ loci to relics**

As we show in the piglet model, very few V $H$  genes are needed and after antigen encounter, further repertoire diversification is by SHM of the same genes (Fig. 10). Furthermore, the

fact that >90% of the repertoire is generated by junctional diversity in HCDR3 (Section 6.2; Fig. 8) and that a transgenic mouse with only one VH is fully immune competent, strengthens the argument. Studies in rabbit also support this view in which one VH gene account for 90% of the early repertoire which can be diversified SHM and SGC after antigen encounter (Winstead et al., 1999; Schiaffella et al., 1999). Detractors from this view may argue that while few (or only one) VH gene are needed, polygeny in the DH and J<sub>H</sub> regions is still necessary (Table 1). Again studies in swine counter this argument since they have only two functional DH segments and one J<sub>H</sub> (Butler et al., 1996; Eguchi-Ogawa et al., 2010). In addition, the chicken has only a single J<sub>H</sub>, DH segments that are nearly identical and only one functional VH and one function V<sub>λ</sub> gene (Fig.3). This species uses the “relics” of upstream pseudogenes for use in repertoire diversification by SGC (Reynaud et al., 1987; Ratcliffe 2006), a mechanism also available to rabbits (Becker & Knight, 1990). Thus we propose that the evolution of gene segment recombination and SHM rendered polygeny in the VH and VL loci of higher vertebrates unnecessary. We believe this polygeny is derived from the tandem array of V-D-J-C clusters in sharks that do not require somatic recombination (Fig. 3). Sharks lack germinal centers as do bony fish and amphibians which makes SHM a less efficient process (DuPasquier et al., 2000). While SHM had been described in these vertebrate classes, we cannot cite head-to-head studies on the frequency of SHM like that presented in Fig. 9B.

We believe that like IgD, the VH, V<sub>λ</sub> and V<sub>κ</sub> polygeny of the most evolved mammals and birds, remain in the genome largely as relics and because of the lack of negative selection. Early vertebrates have as many as four light chain loci but with evolution the number diminished and birds have lost all but lambda. That different Ig isotypes and certain IgG subclass have specific biological functions suggests a selective advantage for polygeny which we believe was true at the dawn of VH and VL duplication that lead to the polygeny in V-region loci. The alternative explanation for especially VH and VL polygeny is that this polygeny occurs in a “hot spot” of RAG-dependent gene segment recombination. For such recombination events to occur, opening of the chromatin structure to provide access to nuclear enzymes, is considered necessary. Perhaps such exposure to repeated recombination activity explains the instability of the locus (Lefranc et al 1983a; Matsuda et al 1990) which renders genes in the locus vulnerable to the molecular machinery involved in gene duplication and genomic gene conversion.

Mainstream immunology has invested almost entirely in studies of mouse and human immune systems and therefore seems to have missed the evolutionary significance of Ig polygeny. To avoid a similar criticism our analysis of polygeny reviewed the major elements of the Ig genes of all vertebrates that have been seriously studied and the mechanisms they use to generate their antibody repertoire. From these comparisons we have offered a hypothesis to explain the polygeny of the major Ig loci in mammals and the reason why at the “highest levels” of vertebrate development, this polygeny appears to be an evolutionary vestige.

## 8. Conclusions

Gene duplication is a common feature of eucaryotic genomes although the degree varies among gene families and loci. It is estimated that ~5% of the human genome is comprised of duplicated genes (Lewin, 2004). Among these are genes of the immunoglobulin supergene family (IGSF). This polygeny is widespread in loci important to the immune system as well as

encoding proteins with only indirect roles in immunity. The  $\beta$ -barrel encoded by genes of the IGSF has obviously been a successful structural motif which can explain its conservation during evolution and diversification into many variants. While IGSF polygeny is widespread, the degree of polygeny is especially pronounced among those that encode the variable heavy and light chain genes of antibodies and the T cell receptor (TCR). Early estimates suggested there were as many as 1000 variable heavy (VH) genes in mice and hundreds in humans. While subsequently studies, including genome projects, have lowered the number of VH genes to ~100-150 in these species, this is still a very large number of homologous genes to occupy a single locus. Similar duplication is seen among genes encoding the variable light chain genes, i.e. V $\lambda$  and V $\kappa$ . However, there are large variations in the numbers and features of duplicated VH and VL genes among mammals and other vertebrates.

This article surveys the duplicated Ig genes in a number of species and uses examples indicating that Ig polygeny resulted from a combination of duplication and genomic gene conversion. Since understanding the evolutionary forces at work in this process requires some understanding of the role played by these duplicated genes in humoral immunity, we review the processes involved in the generation of the antibody repertoire such as Ig gene segment recombination, junctional diversity, somatic hypermutation (SHM) and somatic gene conversion (SGC). We review these processes in various vertebrates but focus on data obtained using the neonatal and newborn piglet model to suggest that evolutionary improvements in somatic processes have reduced the need for the Ig polygeny that evolved among lower vertebrates. We also describe the more recent duplication of the C $\gamma$  genes of mammals that indicates the process was similar. C $\gamma$  genes encode the subclasses of mammalian IgG, the “flagship mammalian antibody” that is unique to this vertebrate class. Since this duplication event occurred more recently, we thought it could provide insight into the advantages conferred by gene duplication.

We propose that the extensive polygeny of VH, V $\lambda$  and V $\kappa$  genes among vertebrates gave adaptive advantage to the earliest vertebrates for generating a diverse repertoire of antibody specificities much as the more recent evolutionary diversification of C $\gamma$  genes resulted in IgG subclass antibodies with different effector functions. We suggest that the evolutionary appearance of mechanisms to somatically alter V-region genes reduced the importance of polygeny in V-region loci for certain mammals and birds. In higher mammals these mechanisms make it possible for a complete functional repertoire to be generated using just one or a few VH genes. This hypothesis can explain why so many of the V-region genes of higher mammals are seldom used, and why deletions of VH genes and C $\gamma$  genes have no effect. We propose that these genes remain as evolutionary vestiges or redundant back-ups in the genome in a manner that parallels the retention of IgD in most mammals. An alternative hypothesis is that the extensive somatic recombination which characterizes the variable region loci (Section 4) creates instability that promotes duplication and genomic gene conversion. In any case, these hypotheses challenge the existing paradigm that random VH, DH and JH recombination among the many gene segments is necessary for survival (presented in immunology textbooks) by placing Ig polygeny into evolutionary perspective.

## 9. Acknowledgement

The authors acknowledge the Molecular Cell Biology Program of the National Science Foundation (USA) and Biological Mimetics of Fredrick, Md for their support of the studies described.

## 10. References

- Amit, A.G., R. A. Mariuzza, S. E. Phillips, & R. J. Pliak. (1986). Three -dimensional structure of an antigen-antibody complex at 2.8 Å resolution. *Science* 233 pp. 747-753.
- Arakawa, H., Furusawa S., Ekino, S., & Yamaguchi, H. (1996). Immunoglobulin gene hyperconversion ongoing in chicken splenic germinal centers. *EMBO J.* 15 pp. 2540-2546.
- Arpin, C., de Bouteillier, O., Razanajaona, D., Briere, F., Banchereau, J., Lecque, S. & Liu, Y. J. (1997). Human peripheral B- cell development. sIgM-IgD + CD38+ hypermutated germinal center centroblasts preferentially express lambda light chains and have undergone mu-to-delta switch. *Ann. N.Y. Acad. Sci.* 815 pp. 193-198.
- Becker, R.S. & Knight, K.L. (1990). Somatic diversification of immunoglobulin heavy chain VDJ genes: evidence for somatic gene conversion in rabbit. *Cell* 63 pp. 987-997.
- Bengten, E., Quiniou, S., Hikima, J., Waldbleser, G., Warr, G.W., Miller, N.W. & Wilson, M. (2006). Structure of the catfish IGH locus: analysis of the region including the single functional IGHM gene. *Immunogenetics* 58 pp. 831-844.
- Bengten, E., S.M. Quiniou, T.B. Stuge, T. Katagirai, N.W. Miller, L.W. Clem, G.W. Warr and M. Wilson. (2002). The IgH locus of the channel catfish, *Ictalurus punctatus*, contains multiple constant region gene sequences: different genes encode heavy chains of membrane and secreted IgD. *J. Immunol.* 169: 2488-2497.
- Blumbach, B., Diehl-Seifer, B., Seack, J., Steffen, R., Muller, I.M. & Muller, W.E.G. (1999). Cloning and expression of new receptors belonging to the immunoglobulin superfamily from the marine sponge *Geodia cydonium*. *Immunogenetics* 49 pp. 751-763.
- Boffey, J., Nicholl, D., Wagner, E.R., Townson, K., Goodyear, C., Furukawa, K., Furukawa, K., Conner, J. & Willison, H.J. (2004). Innate murine B cells produce anti-disialosyl antibodies reactive with *Camylobacter jejuni* LPS and gangliosides that are polyreactive and encoded by a restricted set of unmutated V genes. *J. Neuroimmunol.* 152 pp. 98-111.
- Bratsch, S., Wertz, N., Chaloner, K., Kunz, T. H. & Butler, J. E. (2011). The little brown bat displays a highly diverse V<sub>H</sub>, D<sub>H</sub> and J<sub>H</sub> repertoire but little evidence of somatic hypermutation. *Develop. Comp. Immunol.* 35 pp. 421-430.
- Brezinschek, H-P, Foster, S. J., Brezinschek, R. I., Doerner, T., Domiati-Saad, R. & Lipsky P. E. (1997). Analysis of the human V<sub>H</sub> gene repertoire. *J. Clin. Invest.* 99 pp. 2488-2501.
- Brown, W. R., Kacskovics, I., Amendt, B. Shinde, R., Blackmore, N., Rothschild, M. & Butler, J.E. (1995). The hinge deletion variant of porcine IgA results from a mutation at the splice acceptor site in the first C<sub>α</sub> intron. *J. Immunol.* 154 pp. 3836-3842.
- Butler, J. E. (1998). Immunoglobulins and immune cells in animal milks. In: *Mucosal Immunology*, Ogra, P. L., Mestecky, J., Lamm, M. E., Strober, W., McGhee, J. R. & Bienenstock, J. (Eds.), Chapter 98. Academic Press, New York pp. 1531-1554.
- Butler, J.E., Santiago-K. Mateo, K., Sun, X-Z, Wertz, N., Sinkora, M., Harvey, R. & Francis, D.L. (2011a). Antibody repertoire development in fetal and neonatal piglets. XX. The ileal Peyer's patches are not a site of B cell lymphogenesis and are not required for systemic B cell proliferation and Ig synthesis.

- Butler, J. E., (1974). Immunoglobulins of the mammary secretions. In: *Lactation, a Comprehensive Treatise*, (Larson, B. L. & Smith, V., eds.), Vol. III, Chapter V, pp. 217-255. Academic Press, New York.
- Butler, J. E., & Kehrle Jr., M. E. (2005). Immunocytes and immunoglobulins in milk. In: *Mucosal Immunology*, (J. Mestecky, M. E. Lamm, W. Strober, J. R. McGhee, L. Mayer and J. Bienenstock, eds.), 3rd Edition, Academic Press, NY. pp. 1763-1793.
- Butler, J. E., Francis, D., Freeling, J., Weber, P., Sun, J. & Krieg, A. M. (2005). Antibody repertoire development in fetal and neonatal piglets. IX. Three PAMPs act synergistically to allow germfree piglets to respond to TI-2 and TD antigens. *J. Immunol.* 175 pp. 6772-6785.
- Butler, J. E., Sun, J. & Navarro, P. (1996). The swine immunoglobulin heavy chain locus has a single J<sub>H</sub> and no identifiable IgD. *International Immunology* 8 pp. 1897-1904.
- Butler, J. E., Weber, P., Sinkora, M., Baker, D., Schoenherr, A., Mayer, B. & Francis, D. (2002). Antibody repertoire development in fetal and neonatal piglets. VIII. Colonization is required for newborn piglets to make serum antibodies to T-dependent and type 2 T-independent antigens. *J. Immunol.* 169 pp. 6822-6830.
- Butler, J. E., Weber, P., Sinkora, M., Sun, J., Ford, S. J. & Christenson, R. (2000). Antibody repertoire development in fetal and neonatal piglets. II. Characterization of heavy chain CDR3 diversity in the developing fetus. *J. Immunol.* 165 pp. 6999-7011.
- Butler, J. E. & Sinkora, M. (2007). The isolator piglet: A model for studying the development of adaptive immunity. *Immunol. Res.* 39 pp. 33-51.
- Butler, J. E., Sun, X-Z, Wertz, N., Lager, K. M., Urban Jr., J., Nara, P. & Tobin, G. (2011b). Antibody repertoire development in fetal and neonatal piglets. XXI. VH usage remains constant during development in fetal piglets and postnatally in pigs exposed to environmental antigen.
- Butler, J. E., Zhao, Y., Sinkora, M., Wertz, N. & Kacskovics, I. (2009a). Immunoglobulins, antibody repertoire and B cell development. *Devel. Comp. Immunol.* 33 pp. 321-333.
- Butler, J. E., Lager, K. M., Splichal, I., Francis, D., Kacskovics, I., Sinkora, M., Wertz, N., Sun, J., Zhao, Y., Brown, W. R., DeWald, R., Dierks, S., Muyldermans, S., Lunney, J.K., McCray, P. B., Rogers, C. S., Welsh, M. J., Navarro, P., Klobasa, F., Habe, F. & Ramsoondar, J. (2009b). The Piglet as a Model for B cell and Immune System Development. *Vet. Immunol. Immunopath.* 128 pp. 147-170.
- Butler, J. E., Wertz, N., Zhao, Y., Kunz, T. H., Bratsch, S., Whitaker, J. & Schountz, T. (2010). Two suborders of bats have the canonical isotypes repertoire of other eutherian mammals. *Devel. Com. Immunol.* 35 pp. 272-284.
- Butler, J. E., Weber, P. & Wertz, N. (2006). Antibody repertoire development in fetal and neonatal pigs. XIII. "Hybrid VH genes" and the pre-immune repertoire revisited. *J. Immunol.* 177 pp. 5459-5470.
- Cerato, E., Birkle, S., Portoukalian, J., Mezazigh, A., Chatal, J.F. & Aubry, J. (1997). Variable region gene segments of nine monoclonal antibodies specific to disialgangliosides (GD2, GD3) and their O-acetylated derivatives. *Hybridoma* 16 pp. 307-316.
- Chen, K., Xu, W., Wilson, M., He, B., Miller, N. W., Begnten, E., Edholm, E-S, Santini, P.A., Rath, R., Chiu, A., Cattalinei, M., Litzman, J., Busseel, J., Huang, B., Meini, A., Riesbaeck, K., Cunningham-Rudles, C., Plebani, A. & Cerutti, A. (2009). Immunoglobulin D enhances immune surveillance by activating antimicrobial, proinflammatory and B cell-stimulating programs in basophils. *Nat. Immunol.* 10 pp. 889-898.

- Currier, S. J., Gallara, J. L. & Knight K. L. (1998). Partial genetic map of the rabbit VH chromosomal region. *J. Immunol.* 140 pp. 1651-1659.
- Davies, J. & Riechmann, L. (1995). Antibody VH domains as small recognition units. *Biotechnol.* 13 pp. 475-479.
- Dawicki, W. & Marshall, J.S. (2007). New and emerging roles for mast cells in host defense. *Curr. Opin. Immunol.* 19 pp. 31-18.
- De Genst, E., Saerens, D., Muyldermans, S. & Conrath, K. (2006). Antibody repertoire development in camelids. *Dev. Comp. Immunol.* 30 pp. 187-198.
- deWildt, R. M., Hoet, R. M., van Venrooig, W. J., Tomlinson, I. M. & Winter, G. (1999). Analysis of heavy and light chain pairing indicates that receptor editing shapes the human antibody repertoire. *J. Mol. Bio.* 285 pp. 895-901.
- Diaz, M., Stanfield, R.L., Greenberg, S. & Flajnik, M. F. (2002). Structural analysis, selection and ontogeny of the shark new antigen receptor (IgNAR): identification of a new locus preferentially expressed in early development. *Immunogenetics* 54: 501-512.
- Dildrop, R., Krawinkel, U., Winter, E. & Rajewsky, K. (1985). VH gene expression in murine lipopolysaccharide blasts distribute over the nine known VH-groups and may be random. *Eur. J. Immunology* 15 pp. 1154-1156.
- Dooley, H., Flajnik, M.F. & Porter, J. (2003). Selection and characterization of naturally-occurring single domain (IgNAR) antibody fragments from immunized sharks by phage display. *Mol. Immunol.* 40 pp. 25-33.
- Dooley, H. & Flajnik, M.F. (2006). Antibody repertoire development in cartilaginous fish. *Devel. Comp. Immunol.* 30 pp. 43-56.
- Du Pasquier, L., Robert, J., Courtet, M., & Musmann, R. (2000). B cell development in the amphibian *Xenopus*. *Immunol. Rev.* 175 pp. 201-213.
- Edholm, E-S, Bengton, E., Staffor, J. L., Sahoo, M., Taylor, E. R., Miller, N. W. & Wilson, M. (2010). Identification of two IgD+ B cell populations in channel catfish, *Ictalurus punctatus*. *J. Immunol.* 185 pp. 4082-4094.
- Eguchi-Ogawa, T, Sun, X-Z., Wertz, N., Uenishi, H., Puimi, F., Chardon, P., Wells, K., Tobin, G. J. & Butler, J. E. (2010). Antibody repertoire development in fetal and neonatal piglets. XI. The relationship of VDJ usage and the genomic organization of the variable heavy chain locus. *J. Immunol.* 184 pp. 3734-3742.
- Flanagan, J.G. & Rabbitts, T. H. (1982) Arrangement of human immunoglobulin heavy chain constant region genes implies evolutionary duplication of a segment containing  $\gamma$ ,  $\epsilon$ ,  $\gamma$  and  $\alpha$  genes. *Nature* 300 pp. 709-713.
- Foster, S. J., Brezinschek, H. P., Brezinschek, R. I. & Lipsky, P. E. (1997). Molecular mechanism and selective influences that shape the kappa gene repertoire of IgM+ B cells. *J. Clin. Invest.* 99 pp. 1614-1627.
- Gay, D., Saunders, T., Camper, S. & Weigert, M. (1993). Receptor editing: an approach by autoreactive B cells to escape tolerance. *J. Exp. Med.* 177 pp. 999-1008.
- Glas, A. M., van Monfort, E. H. N. & Milner, E. C. B. (2000). The human antibody repertoire: Old notions, current realities and VH gene-dependent biases. In: *The antibodies*, (Zanetti, M. & Capra, J.D., eds). Vol 6, pp 63-79. Harwood Academic Publishers, Amsterdam.
- Gu, H., Tarlinton, D., Muller, W., Rajewsky, K. & Forster, I. (1991). Most peripheral B cells in mice are ligand restricted. *J. Exp. Med.* 173 pp. 1357-1371

- Hamers-Casterman, C., Atarhouch, T., Muylermans, S., Robinson, G., Hamers, C., Songa, E. B., Bendahman, N. & Hammers, R. (1993). Naturally occurring antibodies devoid of light chains. *Nature* 363 pp. 446-448.
- Hammarstrom L., Lefranc G., Lefranc M-P, Persson, M.A.A. & Smith, C.I.E. (1986). *Monogr. Allergy* 20 pp. 18-25.
- Herrin, B. R. & Cooper, M. D. (2010). Alternative adaptive immunity in jawless vertebrates. *J. Immunol.* 185 pp. 1367-1374.
- Hordvik, I., Thevarajan, J., Sandal, I., Bastani, N. & Krossoy, B. (1999). Molecular cloning and phylogenetics analysis of the Atlantic salmon immunoglobulin D gene. *Scand. J. Immunol.* 50 pp. 202-210.
- Ichiyoshi, Y. & Casali, P. (1994). Analysis of the structural correlates for antibody polyreactivity by multiple reassortments of chimeric human immunoglobulin heavy and light chain V segments *J. Exp. Med.* 180 pp. 885-895.
- Janeway, C.A., Travers, P., Walport, M., & Shlomchik, M.J. (2005). Immunobiology, 6<sup>th</sup> Edition, Garland Science, New York.
- Janssens, R., Dekker, S., Hendriks, R. W., Panayotou, G., van Remoortere, A., San, J. K., Groveld, F. & Drabek, D. (2006). Generation of heavy-chain-only antibodies in mice. *Proc. Nat'l. Acad. Sci. USA* 103 pp. 15130-15135.
- Johnston, C. M., Wood, A. L., Bolland, D. J. & Corcoran, A. E. (2006). Complete sequence assembly and characterization of the C57BL/6 mouse heavy chain V region. *J. Immunol.* 176 pp. 4221-4234.
- Karasuyama, H., Mukai, K., Tsujimura, Y., & Obata, K. (2009). Newly discovered roles for basophils : a neglected minority gains new respect. *Nature Rev. Immunol.* 9 pp. 9-13.
- Keyeux, G, Lefranc, G. & Lefranc, M. P. (1989). A multigene deletion in the human IGH constant region locus involves highly homologous hot spots of recombination. *Genomics* 5 pp. 432-441.
- Klobasa, F., Werhahn, E., & Butler, J.E. (1981). Regulation of humoral immunity in the piglet by immunoglobulins of maternal origin. *Res. Vet. Sci.* 31 pp. 195-206.
- Knight, K. L. (1992). Restricted VH usage and generation of antibody diversity in rabbit. *Ann. Rev. Immunol.* 10 pp. 593-616.
- Koelsch, K., Zheng, N. Y., Zhang, Q., Duty, A., Helms, C., Mathias, M. D., Jared, M., Smith, K., Capra, J. D., & Wilson, P. C. (2007). Mature B cells class switch to IgD are autoreactive in healthy individuals. *J. Clin Invest.* 117 pp. 1558-1565.
- Kranz, D. M. & Voss, Jr., E. W. (1981). Restricted reassociation of heavy and light chains from hapten-specific monoclonal antibodies. *Proc. Nat'l Acad. Sci.* 78 pp. 5807-5811.
- Lavoie, T. B., Mohan, S., Lipschultz, C. A., Grivel, J-C., Li, Y. I., Mainhart, C. R., Kam-Morgan, L. N. W., Drohan, W. N. & Smith-Gill, S. J. (1999). Structural differences among monoclonal antibodies with distinct fine specificities and kinetic properties. *Mol. Immunol.* 36 pp. 1189-1205.
- Lefranc, M. P., Lefranc, G., de Lange, G., Out, T. A., van den Broek, P. J., van Nieuwkoop, J., Radl, J., Hela, A. N., Chaabani, H., van Loghem, E. & Rabbitts, T. H. (1983a) Instability of the human immunoglobulin heavy chain constant region locus indicated by different inherited chromosomal deletions. *Mol. Biol Med.* 1 pp. 207-217.
- Lefranc, G., Chaabani, H., van Loghem, E., Lefranc, M. P., de Lange, G. & Helal, A. N. (1983b). Simultaneous absence of the human IgG1, IgG2, IgG4 and IgA1



- subclasses: immunological and immunogenetical considerations. *Eur. J. Immunol* 13 pp. 240-244.
- Lefranc, M. P., Lefranc, G. & Rabbitts, T. H. (1982). Inherited deletion of immunoglobulin heavy chain constant region genes in normal individuals. *Nature* 300 pp. 760-762.
- Lewin, B., (2004). *Genes VIII*. Pearson-Prentice Hall, Upper Saddle River, N.J. p. 87.
- Liu, Y.J., de Bouteiller, O., Arpin, C., Briere, F., Gailbert, L., Ho, S. , Martinez-Valdez, H., Banchereau, J. & Lebecque, S. (1990). Normal IgD+IgM- germinal center B cells can express up to 80 mutations in the variable region of their IgD transcripts. *Immunity* 4 pp. 603-613.
- Lutz, C., Ledermann, B., Kosco-Vilbois, M. H., Ochsenbein, A. F., Zingernagel, R. M., Kohler, G. & Brombacher, F. (1998). IgD can largely substitute for loss of IgM function in B cells. *Nature* 393 pp. 797-801.
- Mageed, R. A., Marmer, I. J. , Wynn, S. L., Moyes, S. P, Maziak, B. B., Bruggemann, M. & MacKworth-Young, C. G. (2001). Rearrangement of the human heavy chain variable region gene V3-23 in transgenic mice generates antibodies reactive with a range of antigens on the basis of VHCDR3 and residues intrinsic to the heavy chain variable region. *Clin. Exp. Immunol.* 123 pp. 1-8.
- Maki, R., Roeder W., Trawnecker, A., Sidman, C., Wabl, M., Raschke, W. & Tonegawa, S. (1981). The role of DNA rearrangement and alternative RNA processing in the expression of immunoglobulin delta. *Cell.* 24 pp. 353-365
- Malynn, B. A., Berman, J. E., Yancopoulos, G. D., Bona. C. A. & Alt, F. W. (1987). Expression of the immunoglobulin heavy-chain variable gene repertoire. *Curr. Top. Microbiol. Immunol.* pp. 135: 75-94.
- Marchalonis, J. J., Schluter, S. F., Bernstein, R. M. & Edmundson, A. B. (1998). Phylogenetic emergence and molecular evolution of the immunoglobulin family. *Adv. Immunol.* 70 pp. 417-506.
- Marr, S., Morales, H., Bottaro, A., Cooper, M., Flajnik, M. & Robert, J. (2007). Localization and differential expression of activation-induced cytidine deaminase in the amphibian *Xenopus* upon antigen stimulation and during early development. *J. Immunol.* 179 pp. 6783-6789.
- Matsuda, F., Shin, E. K., Nagaoka, H., Matsumura, R., Haino, M., Fukita, Y., Taka-ishi, S., Imai, T., Riley, J. H., Amand, R., Soeda, E. & Honjo, T. (1993). Structure and physical map of 64 variable segments in 3' 0.8-megabase region of the human immunoglobulin heavy-chain locus. *Nature Genet.* 3 pp. 88-94.
- Matsuda, F., Ishii, K., Bourvagnet, P., Kuma, K., Hayashida, H., Miyata, T. & Honjo, T. (1998). The complete nucleotide sequence of the human immunoglobulin heavy chain region locus. *J. Expt. Med.* 188 pp. 2151-2161.
- Matsuda, F., Sin, E. K., Hirabayashi, Y., Nagaoka, H., Yoshida, M. C., Zong, S. Q. & Honjo, T. (1990). Organization of variable region segments of the human immunoglobulin heavy chain: duplication of the D5 cluster within the locus and interchromosomal translocation of variable region segments. *EMBO J.* 9: pp. 2501-2506.
- Meselson, M. S. & Radding, C. M. (1975). A general model for genetic recombination. *Proc. Nat'l Acad. Sci. USA* 72 pp. 358-361.
- Migone, N., Oliviero, S., de Lange, G., Delacroix, D. L., Boschis, D., Altruda, F., Silengo, L., Demarchi, M., & Carbonaro, A. O. (1984). Multiple gene deletions within the human heavy-chain cluster. *Proc. Nat'l Acad. Sci. USA* 81 pp. 5811-5815.

- Miller, M. A., & Steele, R.E. (2000). Lemon encodes an unusual receptor protein-tyrosine kinase expressed during gametogenesis in Hydra. *Dev. Biol.* 224:286-298.
- Mo, J.A., & Holmdahl, R. (1996). The B cell response to autologous type II collagen: biased V gene repertoire with V gene sharing and epitope shift. *J. Immunol.* 157 pp. 2440-2448.
- Monroe, J. G., Havran, W. L., & Cambier, J. C. (1983). B lymphocyte activation entry into cell cycle is accompanied by decrease expression of IgD but not IgM. *Eur. J. Immunol.* 13 pp. 208-213.
- Muyldermans, S., Ghassabeh, G.H., & Saerens D. (2009). Single-domain antibodies. In : *Recombinant antibodies for immunotherapy*, Little, M. (ed.), Cambridge University Press. Chapter 16, pp. 216-230.
- Navarro, P., Christenson, R. Ekhardt, G, Lunney, J.K., Rothschild, M., Bosworth, B, Lemke, J. & Butler, J.E. (2000). Genetic differences in the frequency of the hinge variants of porcine IgA is breed dependent. *Vet. Immunol. Immunopath.* 73 pp. 287-295.
- Nguyen, V., Su, C., Muyldermans, S. van der Loo, W. (2002). Heavy chain antibodies in Camelidae; a case of evolutionary innovation. *Immunogenetics* 54 pp. 39-47.
- Nitschke, L., Kosco, M. L., Kohler, G., & Lamers, M. C. (1993). Immunoglobulin D deficient mice can mount normal immune responses to thymus -independent and -dependent antigens. *Proc. Nat'l Acad. Sci. USA* 90 pp.1887-1891.
- Notarangelo, L. D., Fischer, A., Geha, R.S., Casanova, J-L., Chapel, H, Conley, M.E., Cunningham-Rundles, C., Etzioni, A., Hammarstrom, L., Nonoyama, S., Ochs, H.D., Puck, J., Roifman, C., Seger, R., & Wedgwood, J. (2009). Primary immunodeficiencies: 2009 update. *J. Allergy Clin. Immunol.* 124: 1161-1178.
- Ogawa, K., Wakayama, A., Kunsisada, T., Oril, H., Watanabe, K. & Agata, K. (1998). Identification of a receptor tyrosine kinase involved in germ cell differentiation in planarians. *Biochem. Biophys. Res. Commun.* 248 pp. 204-209.
- Ohta, Y. & Flajnik, M. (2006). IgD, like IgM, is a primordial immunoglobulin class perpetuated in most jawed vertebrates. *Proc. Nat'l. Acad. Sci. USA* 103 pp. 10723-10728.
- Olsson, P.G., Rabbani, H., Hammarstrom, L. & Smith, C.I.E. (1993). Novel human immunoglobulin heavy chain constant region gene deletion haplotypes characterized by pulsed -field electrophoresis. *Clin. Exp. Immunol.* 94: pp. 84-90.
- Padlan, E. A. (1994). Anatomy of the antibody molecule. *Mol. Immunol.* 31 pp. 169-217.
- Plaut, A.G., Wustar, R. Jr. & Capra, J.D. (1974). Differential susceptibility of human IgA immunoglobulins to streptococcal IgA proteases. *J. Clin. Invest.* 54 pp. 1295-1300.
- Rabbani, H., Kondo, N., Smith, C.I., Hammarstrom, L. (1995). The influence of gene deletion and duplication within the IGHC locus on serum immunoglobulin subclass levels. *Clin. Immunol. Immunopathol.* 76:S pp. 214-218.
- Ratcliffe, M. J. H. (2006). Antibodies, immunoglobulin genes and the bursa of Fabricius in chicken. B cell development. *Devel. Comp. Immunol.* 30 pp. 101-118.
- Reiter, Y., Schuck, P., Boyd, L. F. & Plaxin, D. (1999). An antibody single -domain phage display library of a native heavy chain variable region: Isolation of functional single-domain VH molecules with a unique interface. *J. Mol. Bio.* 290 pp. 685-698.
- Retter, I., Chevillard, C., Scharfe, M., Conrad, A., Hafner, M., Im, T-H, Ludewig, M., Nordsied, G., Severitt, S., Thies, S., Mauhar, A., Bloecker, H., Mueller, W. & Riblet, R. (2007). Sequence and characterization of the Ig heavy chain constant and partial variable region of the mouse strain 129S1. *J. Immunol.* 179 pp. 2419-2427.

- Reynaud, C.A., Anquez, V., Daher, A. & Weill, J. (1987). A hyperconversion mechanism generates the chicken pre-immune light chain repertoire. *Cell*. 48 pp. 379-388.
- Rodkey, L. S. & Adler, F. L. (1983). Regulation of natural anti- allotypic antibody responses by network induced auto-anti-idiotypic responsiveness of their offspring. *J. Exp. Med.* 152 pp. 1024-1035.
- Roes, J. & Rajewsky, K. (1993). Immunoglobulin D (IgD)-deficient mice reveal an auxiliary receptor function for IgD in antigen-mediated recruitment of B cells. *J. Exp. Med.* 177 pp. 45-55.
- Rowe, D. S. & Fahey, J. L. (1965). A new class of human immunoglobulins. *J. Expt. Med.* 121 pp. 171-199.
- Rumfelt, L. L., McKinney, E. C., Taylor, E. & Flajnik, M. F. (2002). The development of primary and secondary lymphoid tissues In the nurse shark *Ginglymostoma cirratum*. B cell zones precede dendritic cell immigration and T cell cell zones formation during ontogeny of the spleen. *Scand. J. Immunol.* 56 pp. 130-148.
- Schiaffella, E., Sehgal, D., Anderson, A. O. & Mage, R. G. (1999). Gene conversion and hypermutation during diversification of VH sequences in developing germinal centers of immunized rabbits. *J. Immunol.* 162 pp. 3984-3995.
- Schroeder, H. W. Jr., Hillson, H. L. & Perlmutter, R. M. (1987). Early restriction of human antibody repertoire. *Science* 238 pp. 791-793.
- Schroeder, H.W. Jr, Hillson, H.L. & Perlmutter, R.M. (1990). Structure and evolution of mammalian VH families. *Int'l Immunol* 2 pp. 41-45.
- Sheehan, K. M., Mainville, C. A., Willert, S. & Brodeur, P. H. (1993). The utilization of individual VH exons in the primary repertoire of adult BALB/c mice. *J. Immunol.* 151 pp. 5363-5375.
- Solem, S.T. & Stenvik, J. (2006). Antibody repertoire development in teleosts--a review with emphasis on salmonids and *Gadus morrhua*. *L. Develop. Comp. Immunology* 30 pp. 57-76.
- Spieker-Polet, H., Yam, P.-C., & Knight K. L. (1993). Differential expression of 13 IgA-heavy chain genes in rabbit lymphoid tissues. *J. Immunol.* 150 pp. 5457-5465.
- Srisapoom, P., Ohira, T., Hirona, I. & Aoki, T. (2004). Genes of the constant regions of functional heavy chain of Japanese flounder. *Parealichthys olivaceus*. *Immunogenetics* 56 pp. 292-300.
- Stenvik, J. & Jorgensen, T. O. (2000). Immunoglobulin D (IgD) of Atlantic cod has a unique structure. *Immunogenetics* 51 pp. 452-461.
- Sun, J., Hayward, C., Shinde, R., Christenson, R., Ford, S.P. & Butler, J.E. (1998). Antibody repertoire development in fetal and neonatal piglets. I. Four VH genes account for 80% of VH usage during 84 days of fetal life. *J. Immunol.* 161 pp. 5070-5078.
- Sun, J., Kacsokovics, I., Brown, W. R. & Butler, J. E. (1994). Expressed swine VH genes belong to a small VH gene family homologous to human VH III. *J. Immunol.* 153 pp. 5618-5627.
- Szostak, J. W., Orr-Weaver, T. L. & Rothstein, R. J. (1983). The double-strand break repair model for recombination. *Cell*. 33 pp. 25-35.
- Thomson, C. A., Little, K. Q., Reason, D. C. & Schrader, J. W. (2011). Somatic diversity in CDR3 loops allows single V-genes to encode innate immunological memories for multiple pathogens. *J. Immunol.* 186: pp. 2291-2298.
- Tiegs, S. L., Russell, D. M. & Nemazee, D. (1993). Receptor editing in self-reactive bone marrow B cells. *J. Exp. Med.* 177 pp. 1009-1020.

- Vigliano, F. A., Bemudez, R., Quiroga, M. I. & Nieto, J. M. (2006). Evidence for melanomacrophage centers of teleosts as evolutionary precursors of germinal centers of higher vertebrates: An immunohistochemical study. *Fish and Shellfish Immunology* 21 pp. 467-471.
- Vitetta, E. S., Cambier, J. C., Ligler, F. S., Kettman, J. R. & Uhr, J. W. (1977). B cell tolerance. IV. Differential role of surface IgM and IgD in determining tolerance susceptibility of murine B cells. *J. Exp. Med.* 146 pp. 1804-1808.
- Wang, H. & Shlomchik, M. J. (1998). Maternal Ig mediates neonatal tolerance in rheumatoid factor transgenic mice but tolerance breaks down in adult mice. *J. Immunol.* 160 pp. 2263-2271.
- Wikler, M., Demeur, C., Dewasne, G. & Urbain, J. (1980). Immunoregulatory role of maternal idiotypes. Ontogeny of immune networks. *J. Exp. Med.* 152 pp. 1024-1035.
- Wilson, M., Bengten, E., Miller, N. W., Chen, L. W., Du Pasquier, L. & Warr, G. W. (1997). A novel chimeric Ig heavy chain from a teleost fish shares similarities to IgD. *Proc. Nat'l. Acad. Sci. USA* 94 pp. 4593-4597.
- Winstead, C. R., Zhai, S. K., Sethupathi, P. & Knight, K. L. (1999). Antigen-induced somatic diversification of rabbit IgA genes: gene conversion and point mutation. *J. Immunol* 162 pp. 6602-6612.
- Withers, D. R., Davison, T. F. & Young, J. R. (2005). Developmentally programmed expression of AID in chicken B cells. *Devel. Comp. Immunol.* 29 pp. 651-662.
- Xu, J. L. & Davis, M. M. (2000). Diversity in the CDR3 region of V<sub>H</sub> is sufficient for most antibody specificities. *Immunity* 13 pp. 37-45.
- Yamaguchi, N., Shimizu, S., Hara, T. & Saito, T. (1983). The effector maternal antigenic stimulation upon the active immune responsiveness of their offspring. *Immunology* 50 pp. 229-238.
- Yancopoulos, G. D., Desiderio, S. V., Paskind, M., Kearney, J. F., Baltimore, D. V. & Alt, F. W. (1984). Preferential usage of the most JH proximal VH gene segments in pre-B cell lines. *Nature* 311 pp. 717-733.
- Yasuda, M., Kajiwar, E., Ekino, S., Taura, Y., Hirota, Y., Horiuchi, H., Matsuda, H., Furusawa, S. (2003). Immunobiology of chicken germinal center: I. Changes in surface Ig class expression in the chicken splenic germinal center after antigenic stimulation. *Devel. Comp. Immunol.* 27 pp. 159-166.
- Zhao, Y., Kacskovics, I., Pan-Hammarstrom, Q., Liberles, D. A., Geli, J., Davis, S. K., Rabbani, H. & Hammarstrom, L. (2002). Artiodactyl IgD: the missing link *J. Immunol.* 169 pp. 4408-4416.
- Zhao, Y., Pan-Hammarstrom, Q., Kacskovics, I. & Hammarstrom, L. (2003). The porcine Ig delta gene: unique chimeric splicing of the first constant region domain in its heavy chain transcripts. *J. Immunol.* 171: pp. 1312-1318.
- Zhao, Y., Pan-Hammarstrom, Q., Yu, S., Wertz, N., Zhang, X., Li, N., Butler, J.E. & Hammarstrom, L. (2006). Identification of IgF, a hinge-region containing Ig class and IgD in *Xenopus tropicalis*. *Proc. Natl. Acad. Sci. (USA)* 103 pp. 12087-12092.
- Zheng, N. Y., Wilson, K., Wang, X., Boston, A., Kolar, G., Jackson, S. M., Liu, Y. I., Pascual, V., Capra, J.D. & Wilson, P.C. (2004). Human immunoglobulin selection associated with class switch and possible tolerogenic origin for C delta class-switched B cells. *J. Clin Invest.* 113 pp. 1188-1201.

# Gene Duplication in Insecticide Resistance

Si Hyeock Lee and Deok Ho Kwon

*Department of Agricultural Biotechnology, Seoul National University, Seoul  
Republic of Korea*

## 1. Introduction

Gene duplication is a widely observed phenomenon in all three kingdoms of life and is considered to be a major driving force in the evolution of genomes and organisms. Gene duplication refers to any duplication event of a region of DNA that contains genes, eventually giving rise to gene families. In a classical sense, gene duplication is considered to predate functional diversification. When a duplicated copy is generated, the surplus copy is released from the selective pressure that is posed by random mutations, which allows the rapid accumulation of mutations without deleterious consequences to the organism (Zhang, 2003). The accumulated mutations can increase the fitness of the organism or create a novel function, thereby playing a major role in evolution through functional divergence (Ohno, 1970; Taylor and Raes, 2004). Paralogous gene family members that share a common ancestor gene are generated from a duplication event, which is distinguished from the orthologous genes in different genomes that share a common ancestor as a result of a speciation event (Hurles, 2004). In another theory, instead one copy retains the original function after gene duplication, both of the two copies become to undergo complementary functional diversification, allowing the evolution of an organism over generations (Force et al., 1999). Whole genome duplication events are also common particularly in plant species having polyploidy genomes. Whole genome duplication has influenced the evolutionary path in many species.

One example of extensive gene duplication is the gene amplification. Contrary to gene duplication, which is a doubling mechanism of one gene, gene amplification refers to the process by which the copy number of a particular gene is specifically increased to a greater extent compared to those of other genes, resulting in a dramatic increase in gene dosage. Gene amplification generally results from the repeated replication of a stretch of DNA in a specific region of a genome. Because gene amplification increases the copy number of a gene relatively quickly, it is commonly involved in gene expression control during the development of an organism. Increased copy numbers of a particular gene enables rapid production of a large amount of protein within a short period.

The most common mechanism of gene duplication is homologous recombination by unequal crossing-over between short repeated sequences on homologous segments of chromosomes during meiosis. The replication slippage is also responsible for the duplication of small contiguous repeats of DNA. The possibility and frequency of gene duplication depend on the degree of repetitive sequence distribution between two homologous chromosomes. Detailed information on gene duplication mechanisms can be found elsewhere in this book.

Gene duplication has been implicated as one of insecticide resistance mechanisms. Amplification of detoxification genes such as esterase in the *Culex* mosquito and green peach aphid, *Myzus persicae*, is the first example of a gene duplication (amplification) that is associated with insecticide resistance. These aspects of gene amplification were previously reviewed by Devonshire and Field (Devonshire and Field, 1991). Although the degree is not as extensive as the gene amplification, another example of detoxification gene duplication is found in organophosphate (OP)-resistant sheep blow flies, in which *αE7* gene is duplicated. In this case, the duplication of *αE7* is considered as a way to simultaneously maintain two different resistance alleles. Tandem duplication of two cytochrome P450 (Cyp450) genes was found in a pyrethroid-resistant strain of a major malaria vector mosquito species, suggesting that Cyp450 gene duplication may contribute to pyrethroid resistance by enhancing monooxygenase activity.

Insecticide target gene duplication is involved in insecticide resistance as well. Duplication of the Rdl  $\gamma$ -amino butyric acid (GABA) receptor subunit gene is a possible cause for the resistance to cyclodiene insecticides in *M. persicae*. Dual copies of acetylcholinesterase (AChE) play a role in resistance and adaptation in *Culex* mosquitoes. AChE duplications reduce the fitness cost associated with the mutant *ace* allele in *Culex* mosquitoes. Recently, multiple copies of AChE were confirmed to confer resistance to an OP insecticide in the two-spotted spider mite. Such extensive duplication of AChE provides adaptive advantages in fitness compensation and resistance. Unlike the duplication of detoxification genes, gene dosage control may become a more important issue in the duplication of insecticide target genes because the encoded gene products are directly involved in the transmission of nerve impulses and the maintenance of nervous system homeostasis. Therefore, adaptive compensation for target gene duplication is necessary, particularly when the extent of duplication is severe and linked to a fitness disadvantage.

In this chapter, we reviewed various reported cases of gene duplication involved in insecticide resistance, and we discussed its roles in the resistance evolution and fitness compensation.

## 2. Amplification and duplication of detoxification genes

### 2.1 Esterase gene in *Culex* mosquitoes

The overexpression of two types of esterases, coded at two loci, Est-3 (A esterase) and Est-2 (B esterase), increases esterase activity to confer resistance to OP insecticides in *Culex* mosquito species (Mouches et al., 1986; Mouches et al., 1990; Guillemaud et al., 1997). Multiple overexpressed allozymes have been described as follows: six at the B esterase locus (B1, B2, B4, B5, B6 and B7) and four (A1, A2, A4 and A5) at the A esterase locus (Raymond et al., 1998). In *Culex pipiens quinquefasciatus*, an 800-fold OP resistance is caused by ca. 250-fold amplification of the B1 esterase (*Est-2* locus) (Mouches et al., 1986). Sequence analysis and the characterization of the amplified gene structure has revealed that the amplicon covers at least 30 kb and contains a highly conserved 25-kb "core" carrying a single copy of the esterase gene (2.8 kb) (Mouches et al., 1990). The core is enclosed by two repetitive DNA sequences in other parts of the genome, but not in proximity to the B1 esterase gene, suggesting that the repetitive sequences have a role in the amplification process (Mouches et al., 1990). Analysis of the genomic structure of the Est genes in different OP-resistant strains revealed that two types of genetic alteration mechanisms (transcriptional regulation and gene amplification) are involved in the development of resistance. Overexpression of one A1

esterase reported in the southern French mosquito strain was due to a transcriptional regulatory mechanism, whereas in other cases, the coamplification of the A and B esterase loci or an amplification of the B esterase locus alone resulted in overproduction of esterase (Guillemaud et al., 1997; Raymond et al., 1998). The level of gene amplification varies depending on different esterase alleles. For example, the copy number of B1 esterase reached 250 copies (Mouches et al., 1986), whereas that of B4 esterase did not exceed a few copies (Guillemaud et al., 1997). The level of gene amplification is also different within and between populations as in the case of the coamplified A2-B2 esterases (Callaghan et al., 1998). The number of independent amplification events at the loci of A and B esterases cannot be precisely estimated. Considering the protein quantification profiles and the molecular data published to date, however, the number of independent amplification events may range from five to ten events as a minimum figure (Raymond et al., 1998). This relatively low frequency of independent amplification events on a global scale suggests that the fitness advantage of esterase gene amplification may be restrictive (Raymond et al., 1998).

## **2.2 Esterase gene in the green peach aphid, *Muzus persicae*, and other aphids**

Overexpression of the esterase (E4) responsible for broad insecticide resistance in the green peach aphid, *M. persicae* Sulz, was also associated with amplification of the structural E4 gene or its closely related variant (FE4). The extent of amplification was well correlated with the activity of the esterase and the level of resistance (Field et al., 1988). Molecular studies revealed that the presence of the amplified E4 gene is correlated with an autosomal 1,3 translocation, whereas amplified FE4 genes are found in insecticide-resistant aphids with normal untranslocated chromosomes (Field et al., 1988). Subsequent *in situ* hybridization assays revealed that a single amplified site is located on autosome 3 near the breakpoint of the autosomal 1,3 translocation in all of the E4-producing aphid clones except one having two other E4 loci. In the most resistant aphid clone producing FE4, however, the amplification sites were widely distributed around the genome (from three to eight sites) (Blackman et al., 1995; Blackman et al., 1999). The relative esterase gene copy numbers in aphid clones with different levels of insecticide resistance (R1, R2 and R3) were determined to increase ca. 4-fold between susceptible, R1, R2 and R3 aphids, reaching a maximum increase of approximately 80-fold amplification in R3 aphids. This proportionate correlation between amplification and resistance further suggested that transcriptional upregulation of amplified genes may not be involved in resistance (Field et al., 1999). The amplified esterase genes are arranged as tandem repeats at a single locus in some aphid clones, whereas amplicons are dispersed throughout the genome in other clones. The amplified E4 and FE4 genes are methylated at the CpG repeats within the gene (Hick et al., 1996). However, the methylation is absent from upstream regions, including the 5' CpG-rich region around the regulatory region, and from 3' flanking DNA. Contrary to the common belief that methylation suppresses gene expression, methylated E4 genes are expressed and loss of the 5-methylcytosine is correlated with transcription suppression, suggesting that the methylation of E4 has a positive role in expression (Field, 2000).

## **2.3 Esterase gene in the sheep blow fly, *Lucilia cuprina***

Resistance to diazinon and malathion is primarily due to two point mutations (Gly137Asp and Trp251Leu) in a carboxylesterase gene (*LcaE7*), encoding both OP hydrolase and malathion carboxylesterase (MCE) activities (Campbell et al., 1997; Campbell et al., 1998).

The OP hydrolase activity, which is responsible for the resistance to diazinon and the majority of other OPs, is conferred by Gly137Asp. In contrast, high MCE activity, which is associated with resistance that is limited to a type of OPs with carboxyl ester bonds such as malathion, is conferred by Trp251Leu mutation. Flies showing double resistance to both malathion and diazinon were also found though they are not common. Since the allele containing double mutations of Gly137Asp and Trp251Leu cannot confer resistance to both malathion and diazinon (Heidari et al., 2004), it was proposed that a duplication of the region containing the *LcaE7* gene generated two loci, each carrying different mutation (Newcomb et al., 2005). As in the case of *LcaE7*, when two (or multiple) mutations cannot exit together in a single locus, duplication may be the most useful option to acquire the benefits of both resistance mutations. In this scenario, the *LcaE7* duplication events most likely postdate the origin of the two resistance alleles, representing a case of gene that share preceding gene duplications (Newcomb et al., 2005). The relatively low frequencies of cases indicating double resistance suggests restrictions on the rate of recombination or fitness costs associated with the duplications.

## 2.4 Cyp450 gene in *M. persicae*

Cyp450-mediated detoxification acts as a primary mechanism in neonicotinoid resistance in *M. persicae* (Puinean et al., 2010). Microarray analysis of all known detoxification genes in *M. persicae* revealed constitutive over-expression (22-fold) of a single Cyp450 gene (*Cyp6CY3*). The overexpression of Cyp450 is due, at least in part, to an approximately 9-fold gene amplification, as quantitative PCR of genomic DNA showed that the diploid genome of a susceptible aphid clone carries two copies of the *Cyp6CY3*, whereas the neonicotinoid-resistant clone has 18 copies (Puinean et al., 2010). Transcriptional upregulation based on mutations in *cis*-acting and/or *trans*-acting regulatory loci has been reported to be mainly responsible for the overexpression of Cyp450 genes in insecticide-resistant insects (Li et al., 2007). Therefore, this may be the first case of Cyp450 amplification that is associated with insecticide resistance in an agriculturally important insect pest but the *Cyp6CY3* amplification mechanism remains to be elucidated.

## 3. Duplication of insecticide-target site genes

### 3.1 Duplication of $\gamma$ -aminobutyric acid (GABA) receptor

A point mutation (Ala302Ser or Ala302Gly) in the 'Resistance to dieldrin' (*Rdl*) gene encoding a GABA receptor subunit is known to confer resistance to cyclodiene insecticides in *Drosophila* and other insects (ffrench-Constant et al., 1993; ffrench-Constant et al., 2000). Because *Rdl* is a single copy gene in most insects, individuals can carry only two different alleles. In contrast, *M. persicae* is reported to have up to four different *Rdl*-like alleles (Anthony et al., 1998). Along with the wild-type allele (encoding Ala302 or allele A), three other alleles encoding Gly302 (allele G), Ser302 (encoded from 'TCG' codon; allele S) and Ser302 (encoded from 'AGT' codon; allele S') were found in *M. persicae* (Anthony et al., 1998). Three of these alleles (A, G and S) were common in individual aphids or aphid clones. The presence of two independent *Rdl* loci in *M. persicae* has been confirmed by Southern analysis in conjunction with sequencing downstream of the exon containing the mutation. Sequence comparison between two loci has suggested that the loci may have been generated through a recent gene duplication event. Interestingly, only one locus carrying the opposite



alleles of the alanine vs. glycine was responsible for resistance, whereas the other locus carrying the two serine-containing alleles (S or S') was not associated with resistance (Anthony et al., 1998). Taken together, it appears that, after gene duplication, one locus (S/S') of *Rdl* likely had been fixed in aphids regardless of their resistance status, whereas the other (A/G) locus had been specialized as a mechanism for GABA receptor insensitivity to dieldrin. This is a typical example of the duplication of a gene encoding an insecticide target site followed by the functional diversification with respect to insecticide resistance.

### 3.2 Duplication of acetylcholinesterase (AChE) gene in *Culex* mosquitoes

In *C. pipiens* mosquitoes, the *ace1* locus encodes acetylcholinesterase 1 (AChE1), which is the target of OP and carbamate insecticides. Most OP-resistant populations showed insensitive AChE1, whereas, in two Caribbean populations, individual mosquito displayed a mixture of sensitive and insensitive AChE1. The parsimonious explanation for these phenomenon is the existence of two *ace1* loci encoding both resistant and susceptible AChE1 (phenotype RS), which were most likely generated by gene duplication (Bourguet et al., 1996). The OP-resistant allele, *ace1R*, has been determined to be due to a single amino acid substitution, Gly119Ser (Weill et al., 2003; Weill et al., 2004), which causes not only reduced sensitivity to OP insecticide but also high fitness cost by modifying the catalytic properties of AChE1 (Weill et al., 2003). Later on, a similar duplication event was suggested for *C. pipiens* populations from Southern France, where an excess of the [RS] phenotype was observed in natural populations (Lenormand et al., 1998). The duplication event in Southern France was dated back to 1993, 15 years after *ace1R* was first detected in the area, and then has gradually replaced the *ace1R* allele in treated areas (Lenormand et al., 1998).

Advantages and costs of *ace1* duplications in relation to OP resistance have previously been described from the perspective of gene dosage (Labbé et al., 2007). Gene dosage is very important to maintain essential cellular processes, and increase in gene dosage by duplication likely disturb this balance (Kondrashov et al., 2002; Papp et al., 2003; Veitia, 2005). Duplication of *ace1* produces higher levels of AChE1, resulting in hyperactivity. In the resistant form of AChE1 (AChE1R), however, the catalytic activity is less than 60% of the susceptible form of AChE1 (AChE1S) (Bourguet et al., 1996; Bourguet et al., 1997). Therefore, *ace1* duplication may restore its normal gene dosage, otherwise phenotypically reduced by the resistance mutation (Labbé et al., 2007). A duplication of two *ace1R* copies may partly restore normal AChE1 activity to a level that is comparable to that of a single copy of *ace1S*, serving as a transitional step to *ace1D* haplotypes (having both *ace1S* and *ace1R*). AChE1 activity in *ace1D* homozygotes is similar to or greater than those in susceptible *ace1S* homozygotes (Bourguet et al., 1996). The slightly higher AChE1 activity generated by *ace1D* may result in another type of fitness cost by rapidly degrading the neurotransmitter acetylcholine. In summary, *ace1* duplication generating both resistant and susceptible copies may be selected as a compensatory mechanism for the fitness cost by the homozygous *ace1R* allele (Labbé et al., 2007). The generation of persistent heterozygosis, in which the susceptible *ace1* allele is always expressed, likely reduces the fitness cost associated with the resistant allele. The duplication event was proposed to occur relatively more frequently than generally conceived and very recently (i.e., within less than past 40 years), demonstrating that the gene duplication associated with insecticide resistance is a typical example of rapidly developing evolutionary events (Labbé et al., 2007).

### 3.3 Extensive duplication of AChE gene in the two-spotted spider mite (TSSM)

Three point mutations (Gly228Ser, Ala391Thr and Phe439W) have been identified using extensive sequence comparisons of the AChE gene from the TSSM (*Tuace*) between OP-resistant and susceptible strains. In addition, their functional roles have been assessed by analyzing the correlation between mutation frequencies and actual resistance levels of several field populations (Kwon et al., 2010b). The frequencies of the Gly228Ser and Phe439Trp mutations in resistant strains never reached 100% even after extensive selection with monocrotophos (Kwon et al., 2010b). To determine whether the lack of saturation for these mutation frequencies is due to the heterozygosity of the *Tuace* allele in individual mites, the frequencies of the three mutations in an individual diploid virgin female and her parthenogenetic haploid male progenies were have been determined using quantitative sequencing (Kwon et al., 2010a). The actual frequencies of the G228S, A391T and F439W mutations in a female mite and its haploid male progenies have been estimated as approximately 50%, 100% and 75%, respectively. These findings clearly suggested the presence of multiple copies of *Tuace*. Determination of *Tuace* copy number in three mite strains (highly resistant AD, moderately resistant PyriF and susceptible UD strains) using quantitative PCR has revealed that resistant strains have relatively more *Tuace* copies than the susceptible strain and that the levels of transcript were directly proportional to copy numbers (Kwon et al., 2010a). AChEs from the AD and PyriF strains have shown reduced catalytic efficiencies based on lower  $k_{cat}$  values, suggesting that the resistant form of AChE is likely accompanied by fitness cost. Relative copy numbers of *Tuace* in field populations of TSSM ranged from 2.4 to 6.1 and are highly correlated with the respective resistance level, suggesting that *Tuace* duplication itself contributes to resistance (Kwon et al., 2010a).

Western blot analysis using AChE-specific antibodies has been conducted to determine whether *Tuace* duplication results in TuAChE overproduction. The protein quantities of TuAChE in seven field-collected mite populations precisely correlated with the copy numbers (Lee and Kwon, unpublished data). To investigate the effects of each mutation on AChE insensitivity and possible fitness costs, eight variants of TuAChE were expressed *in vitro* using the baculovirus expression system. Kinetic analysis revealed that the Ala391Thr mutation did not alter the kinetic properties of AChE, whereas the Gly228Ser and Phe439Trp mutations significantly increased the insensitivity to monocrotophos. Moreover, when the Gly228Ser and Phe439Trp mutations were co-expressed, insensitivity increased over 1000-fold. These results show that both mutations confer resistance in a synergistic manner. However, the presence of the mutations considerably reduced the catalytic efficiency of AChE, suggesting an apparent fitness cost in monocrotophos-resistant mites. Reconstitution of the multiple copies of AChE with different compositions of the mutations revealed that the catalytic efficiencies of the six-copy and two-copy AChEs (resembling the AD and PyriF strains of mite, respectively) were lower but comparable to that of wildtype AChE. These finding clearly suggest that multiple rounds of *Tuace* duplication is needed to compensate the reduced catalytic activity of AChE caused by mutations. Whether mutation or gene duplication occurs first is unknown. However, the introduction of a single mutation (Gly228Ser or Phe439Trp) or a double mutation in a single copy of *Tuace* is unlikely because the fitness cost is severe based on the dramatic reductions in the catalytic efficiency. Therefore, at least a single event of *Tuace* duplication predates the introduction of mutations. A single Gly228Ser mutation likely occurs first in one of the duplicates as seen in the PyriF strain, in which the Gly228Ser mutation has been identified in one of the *Tuace* duplicates. Under continuous selection pressure by OPs, further duplication of mutations might have

been required to compensate for the further reduction of catalytic efficiency that is attributed to an additional Phe439Trp mutation. In summary, monocrotophos resistance in TSSM may have evolved through a combination of phased gene duplication and mutation accumulation.

#### 4. Conclusions

Duplications (or amplifications) of resistance-related genes are frequent mechanisms in insecticide resistance. Extensive forms of gene duplication (i.e., amplification) are commonly found in metabolic genes that are involved in insecticide detoxification, including esterase and Cyp450. In these cases, even with dramatically increased gene dosage, the apparent fitness cost is not severe. In other words, low fitness costs that are associated with high dosages of metabolic genes allow the amplification of genes. Gene duplication events have been found in insecticide target genes, including *ace* and *Rdl*, which play crucial functions in nerve impulse transmission. Unlike the amplification of metabolic genes, the level of duplication of these genes is precisely regulated due to the necessity to maintain the normal gene dosage. Mutations conferring target site insensitivity are always accompanied with duplication events. Because mutations in the insecticide target sites frequently alter catalytic or functional properties of target proteins, which usually increase fitness costs, duplication may act as a compensatory mechanism to restore the normal activity of the target protein, which is otherwise detrimental to maintaining the nervous system homeostasis. Conversely, if the increase of the target protein dose due to an incidental gene duplication event has different fitness consequences, the introduction and selection of any target site mutations conferring insecticide resistance is facilitated following duplication because the mutations that are associated with resistance usually reduce the normal function of target proteins. Taken together, it is difficult to determine whether duplication or mutation occurs first. However, these evolutionary events to acquire insecticide resistance may interact each other to balance the level of resistance and accompanied fitness costs. Genetic introgression between different populations plays a crucial role in spreading and formulating the degree of gene duplication and mutation. In *Anopheles gambiae*, for example, the genetic traits of the *ace1R* mutation and the *ace1* duplication are shared between populations through introgression (Djogbénou et al., 2008).

#### 5. References

- Anthony, N., Unruh, T., Ganser, D. & French-Constant, R. (1998). Duplication of the *Rdl* GABA receptor subunit gene in an insecticide-resistant aphid, *Myzus persicae*. *Molecular and General Genetics*, Vol.260, No.2-3, (November 1998), pp. 165-75, ISSN 0026-8925
- Blackman, R. L., Spence, J. M., Field, L. M. & Devonshire, A. L. (1995). Chromosomal location of the amplified esterase genes conferring resistance to insecticides in *Myzus persicae* (Homoptera: Aphididae). *Heredity*, Vol.75, No.3, (September 1995), pp. 297-302, ISSN 0018-067X
- Blackman, R. L., Spence, J. M., Field, L. M. & Devonshire, A. L. (1999). Variation in the chromosomal distribution of amplified esterase (*FE4*) genes in Greek field populations of *Myzus persicae* (Sulzer). *Heredity*, Vol.82, No.2, (February 1999), pp. 180-186, ISSN 0018-067X

- Bourguet, D., Lenormand, T., Guillemaud, T., Marcel, V., Fournier, D. & Raymond, M. (1997). Variation of dominance of newly arisen adaptive genes. *Genetics*, Vol.147, No.3, (November 1997), pp. 1225-1234, ISSN 0016-6731
- Bourguet, D., Raymond, M., Bisset, J., Pasteur, N. & Arpagaus, M. (1996). Duplication of the *Ace1* locus in *Culex pipiens* mosquitoes from the Caribbean. *Biochemical Genetics*, Vol.34, No.9-10, (October 1996), pp. 351-62, ISSN 0006-2928
- Callaghan, A., Guillemaud, T., Makate, N. & Raymond, M. (1998). Polymorphisms and fluctuations in copy number of amplified esterase genes in *Culex pipiens* mosquitoes. *Insect Molecular Biology*, Vol.7, No.3, (August 1998), pp. 295-300, ISSN 1365-2583
- Campbell, P. M., Newcomb, R. D., Russell, R. J. & Oakeshott, J. G. (1998). Two different amino acid substitutions in the ali-esterase, E3, confer alternative types of organophosphorus insecticide resistance in the sheep blowfly, *Lucilia cuprina*. *Insect Biochemistry and Molecular Biology*, Vol.28, No.3, (March 1998), pp. 139-150, ISSN 0965-1748
- Campbell, P. M., Trott, J. F., Claudianos, C., Smyth, K. A., Russell, R. J. & Oakeshott, J. G. (1997). Biochemistry of esterases associated with organophosphate resistance in *Lucilia cuprina* with comparisons to putative orthologues in other Diptera. *Biochemical Genetics*, Vol.35, No.1-2, (February 1997), pp. 17-40, ISSN 0006-2928
- Devonshire, A. L. & Field, L. M. (1991). Gene Amplification and Insecticide Resistance. *Annual Review Entomology*, Vol.36, No.1, (January 1991), pp. 1-21, ISSN 0066-4170
- Djogbénou, L., Chandre, F., Berthomieu, A., Dabiré, R., Koffi, A., Alout, H. & Weill, M. (2008). Evidence of introgression of the *ace-1<sup>R</sup>* mutation and of the *ace-1* duplication in west african *Anopheles gambiae* s. s. *PLoS ONE*, Vol.3, No.5, (May 2008), pp. e2172, ISSN 1932-6203
- ffrench-Constant, R. H., Anthony, N., Aronstein, K., Rocheleau, T. & Stilwell, G. (2000). Cyclodiene insecticide resistance: from molecular to population genetics. *Annual Review Entomology*, Vol.45, No.1, (January 2000), pp. 449-466, ISSN 0066-4170
- ffrench-Constant, R. H., Rocheleau, T. A., Steichen, J. C. & Chalmers, A. E. (1993). A point mutation in a *Drosophila* GABA receptor confers insecticide resistance. *Nature*, Vol.363, No.6428, (June 1993), pp. 449-451, ISSN 1476-4687
- Field, L. M. (2000). Methylation and expression of amplified esterase genes in the aphid *Myzus persicae* (Sulzer). *Biochemical Journal*, Vol.349 Pt 3, (August 2000), pp. 863-8, ISSN 0264-6021
- Field, L. M., Blackman, R. L., Tyler-Smith, C. & Devonshire, A. L. (1999). Relationship between amount of esterase and gene copy number in insecticide-resistant *Myzus persicae* (Sulzer). *Biochemical Journal*, Vol.339, No.3, (May 1999), pp. 737-742, ISSN 0264-6021
- Field, L. M., Devonshire, A. L. & Forde, B. G. (1988). Molecular evidence that insecticide resistance in peach-potato aphids (*Myzus persicae* Sulz.) results from amplification of an esterase gene. *Biochemical Journal*, Vol.251, No.1, (April 1988), pp. 309-312, ISSN 0264-6021
- Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, Vol.151, No.4, (April 1999), pp. 1531-45, ISSN 0016-6731

- Guillemaud, T., Makate, N., Raymond, M., Hirst, B. & Callaghan, A. (1997). Esterase gene amplification in *Culex pipiens*. *Insect Molecular Biology*, Vol.6, No.4, (November 1997), pp. 319-327, ISSN 1365-2583
- Heidari, R., Devonshire, A. L., Campbell, B. E., Bell, K. L., Dorrian, S. J., Oakeshott, J. G. & Russell, R. J. (2004). Hydrolysis of organophosphorus insecticides by in vitro modified carboxylesterase E3 from *Lucilia cuprina*. *Insect Biochemistry and Molecular Biology*, Vol.34, No.4, (April 2004), pp. 353-63, ISSN 0965-1748
- Hick, C. A., Field, L. M. & Devonshire, A. L. (1996). Changes in the methylation of amplified esterase DNA during loss and reselection of insecticide resistance in peach-potato aphids, *Myzus persicae*. *Insect Biochemistry and Molecular Biology*, Vol.26, No.1, (January 1996), pp. 41-47, ISSN 0965-1748
- Hurles, M. (2004). Gene duplication: the genomic trade in spare parts. *PLoS Biology*, Vol.2, No.7, (July 2004), pp. E206, ISSN 1545-7885
- Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biology*, Vol.3, No.2, (January 2002), pp. RESEARCH0008, ISSN 1465-6914
- Kwon, D. H., Clark, J. M. & Lee, S. H. (2010a). Extensive gene duplication of acetylcholinesterase associated with organophosphate resistance in the two-spotted spider mite. *Insect Molecular Biology*, Vol.19, No.2, (April 2010), pp. 195-204, ISSN 1365-2583
- Kwon, D. H., Im, J. S., Ahn, J. J., Lee, J.-H., Marshall Clark, J. & Lee, S. H. (2010b). Acetylcholinesterase point mutations putatively associated with monocrotophos resistance in the two-spotted spider mite. *Pesticide Biochemistry and Physiology*, Vol.96, No.1, (January 2010), pp. 36-42, ISSN 0048-3575
- Labbé, P., Berticat, C., Berthomieu, A., Unal, S., Bernard, C., Weill, M. & Lenormand, T. (2007). Forty years of erratic insecticide resistance evolution in the mosquito *Culex pipiens*. *PLoS Genetics*, Vol.3, No.11, (November 2007), pp. e205, ISSN 1553-7404
- Lenormand, T., Guillemaud, T., Bourguet, D. & Raymond, M. (1998). Evaluating gene flow using selected markers: a case study. *Genetics*, Vol.149, No.3, (July 1998), pp. 1383-1392, ISSN 0016-6731
- Li, X., Schuler, M. A. & Berenbaum, M. R. (2007). Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. *Annual Review Entomology*, Vol.52, (August 2006), pp. 231-53, ISSN 0066-4170
- Mouches, C., Pasteur, N., Berge, J., Hyrien, O., Raymond, M., de Saint Vincent, B., de Silvestri, M. & Georghiou, G. (1986). Amplification of an esterase gene is responsible for insecticide resistance in a California *Culex* mosquito. *Science*, Vol.233, No.4765, (August 1986), pp. 778-780, ISSN 0036-8075
- Mouches, C., Pauplin, Y., Agarwal, M., Lemieux, L., Herzog, M., Abadon, M., Beyssat-Arnaouty, V., Hyrien, O., de Saint Vincent, B. R., Georghiou, G. P. & Pasteur, N. (1990). Characterization of amplification core and esterase B1 gene responsible for insecticide resistance in *Culex*. *Proceedings of the National Academy of Sciences of the United States of America*, Vol.87, No.7, (April 1990), pp. 2574-8, ISSN 0027-8424
- Newcomb, R. D., Gleeson, D. M., Yong, C. G., Russell, R. J. & Oakeshott, J. G. (2005). Multiple mutations and gene duplications conferring organophosphorus insecticide resistance have been selected at the *Rop-1* locus of the sheep blowfly,

- Lucilia cuprina*. *Journal of Molecular Evolution*, Vol.60, No.2, (February 2005), pp. 207-220, ISSN 0022-2844
- Ohno, S. (1970). *Evolution by gene duplication*, Springer-Verlag, ISBN 0-04-575015-7, New York, USA
- Papp, B., Pal, C. & Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, Vol.424, No.6945, (July 2003), pp. 194-7, ISSN 1476-4687
- Puinean, A. M., Foster, S. P., Oliphant, L., Denholm, I., Field, L. M., Millar, N. S., Williamson, M. S. & Bass, C. (2010). Amplification of a cytochrome P450 gene is associated with resistance to neonicotinoid insecticides in the aphid *Myzus persicae*. *PLoS Genetics*, Vol.6, No.6, (June 2010), pp. e1000999, ISSN 1553-7404
- Raymond, M., Chevillon, C., Guillemaud, T., Lenormand, T. & Pasteur, N. (1998). An overview of the evolution of overproduced esterases in the mosquito *Culex pipiens*. *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol.353, No.1376, (October 1998), pp. 1707-11, ISSN 0962-8436
- Taylor, J. S. & Raes, J. (2004). Duplication and divergence: the evolution of new genes and old ideas. *Annual Review Genetics*, Vol.38, No.1, (June 2004), pp. 615-643, ISSN 0066-4197
- Veitia, R. A. (2005). Gene dosage balance: deletions, duplications and dominance. *Trends in Genetics*, Vol.21, No.1, (January 2005), pp. 33-35, ISSN 0168-9525
- Weill, M., Lutfalla, G., Mogensen, K., Chandre, F., Berthomieu, A., Berticat, C., Pasteur, N., Philips, A., Fort, P. & Raymond, M. (2003). Comparative genomics: insecticide resistance in mosquito vectors. *Nature*, Vol.423, No.6936, (May 2003), pp. 136-7, ISSN 1476-4687
- Weill, M., Malcolm, C., Chandre, F., Mogensen, K., Berthomieu, A., Marquine, M. & Raymond, M. (2004). The unique mutation in *ace-1* giving high insecticide resistance is easily detectable in mosquito vectors. *Insect Molecular Biology*, Vol.13, No.1, (February 2004), pp. 1-7, ISSN 1365-2583
- Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, Vol.18, No.6, (June 2003), pp. 292-298, ISSN 0169-5347

# Gene Duplication and the Origin of Translation Factors

Galina Zhouravleva and Stanislav Bondarev  
*Department of Genetics and Breeding, St Petersburg State University*  
*Russia*

## 1. Introduction

Charles Darwin is famous for his contribution to the development of evolutionary theory. Less commonly known is that Darwin was a good botanist. He wrote several books devoted to flowering plants. Being an honest scientist, he did not conceal the inability of his theory of evolution to explain the sudden appearance and rapid spread of angiosperms, calling this phenomenon an “abominable mystery”. One possible solution to the puzzle that agitated Darwin may be the several successive duplications of the ancient ancestral genome at the beginning of the divergence of angiosperms that gave them the ability to rapidly accumulate changes (Cui et al., 2006). Speculation about the possible role of gene duplication in evolution began in the middle of the last century (Sturtevant, 1925; Haldane, 1932; Muller, 1936; Lewis, 1951), but only the later rapid development of molecular biology allowed the identification of numerous repeated sequences that revealed a high frequency of gene duplication in evolution. Based on this information, S. Ohno (Ohno, 1970) suggested that gene duplication was the only way new genes could emerge.

## 2. Types of duplications

Duplication of DNA can occur in many ways: (1) partial duplication of a gene (or an internal duplication), (2) duplication of a single gene, (3) partial duplication of a chromosome, (4) duplication of an entire chromosome, and (5) genome duplication, or polyploidy. The first four types of duplication are sometimes combined under the term SSD (smaller scale duplication) (Davis & Petrov 2005). Other authors prefer the terms “paralogon” (derived from “paralog”), for extended duplicated regions containing paralogs, and SGD (single gene duplication), for duplications of individual genes (Durand & Hoberman, 2006). Duplication of the entire genome is designated as WGD (whole genome duplication) (Davis & Petrov, 2005). According to Ohno, duplication of the genome rather than its individual parts is more important for evolution, because the partial duplication of regulatory genes or other restricted elements of the genome may lead to regulatory imbalances (Ohno 1970).

### 2.1 Whole genome duplications

Ancient polyploidizations of the genome have been identified in all four eukaryotic kingdoms: plants, animals, fungi and protists. In all cases, the proportion of genes in the

form of duplicated copies ranges from 10 to 50% and often correlates with the time elapsed since duplication (Scannell et al., 2006).

WGD is widespread in plants (Vision et al., 2000; Adams & Wendel, 2005). Estimates of the incidence of polyploidy in angiosperms vary from 30 to 80%, and about 3% of speciation events are explained by genome duplications (Otto & Whitton, 2000). Many, if not all, species of plants may thus have at least one polyploid ancestor. Most eudicots are assumed to have an ancient hexaploid ancestor, with subsequent tetraploidization in some taxa (Jaillon et al., 2007).

Duplication of the entire genome in the yeast *Saccharomyces cerevisiae* led to an initial increase in the number of genes from 5000 to 10 000, but the subsequent loss of paralogs has led to the preservation in modern *Saccharomyces* of about 5500 protein-coding genes, of which 1102 form 551 paralogous pairs (Byrne & Wolfe, 2005). A special term, ohnologs, dedicated to S. Ohno, was proposed for paralogs resulting from WGD (Wolfe, 2000).

Detection of natural polyploidy is a difficult task, especially for ancient events. Recent duplications can be detected by comparing closely related species, one of which underwent diploidization and therefore contains twice as many chromosomes as species that did not undergo WGD. For example, a comparison of the genomes of *Ashbya gossypii* and *S. cerevisiae* revealed that both species evolved from a single ancestor that had seven or eight chromosomes (Dietrich et al., 2004). Changes in chromosome number due to mutations (in particular translocations) led to the ancestors of *A. gossypii* and *S. cerevisiae*. WGD in *S. cerevisiae* has provided this species with new opportunities for functional divergence absent in *A. gossypii*. A similar comparative analysis was also carried out for *S. cerevisiae* and its closest non-WGD relative, *Kluyveromyces waltii* (Kellis et al., 2004).

The older the duplication, the harder the analysis, because a period of diploidization often follows polyploidization, which "transforms" the polyploid genome to the diploid state. Diploidization is achieved by an intensive loss of genes, rearrangements of the genome and the divergence of duplicated genes. Recent analyses have also shown that the duplication of individual genes in evolution has occurred much more frequently than was previously thought (Lynch & Conery, 2000; Lynch et al., 2001). Diploidization has been studied in many genomes including those of plants (Chapman et al., 2006; Jaillon et al., 2007; Tuskan et al., 2006), bony fishes (Brunet et al., 2006), yeasts (Piskur, 2001; Kellis et al., 2004; Scannell et al., 2006; Scannell et al., 2007), *Paramecium* (Aury et al., 2006) and vertebrates (Blomme et al., 2006).

Plants have repeatedly undergone polyploidization during evolution, presumably aided by their ability to propagate vegetatively and by the existence of specific regulatory mechanisms in plant cells. In particular, model polyploids have been characterized by a rapid loss of some genes and the specific inactivation of others by methylation (Kashkush et al., 2002; Comai et al., 2000; Lee & Chen, 2001). Epigenetic silencing may protect the duplicated copies from pseudogenization, thus facilitating the acquisition of new functions (Rodin & Riggs, 2003).

Vertebrate genomes contain many families of genes that are not found in invertebrates, and many gene duplications apparently occurred early in the evolution of the chordates (Taylor & Raes, 2004). Ohno suggested that the complex genome of vertebrates arose as a result of two rounds (2R) of WGD (Ohno, 1970). This view was once supported by the belief that the human genome contained about 100 000 genes, which was four times more than the estimated number of genes in the genomes of invertebrates. Sequencing of the human genome has since reduced the estimate of the number of genes to 20 000-25 000 but has not



yet answered the question of the number of duplications of the ancestral genome. Some authors continue to support the 2R hypothesis (Larhammar et al., 2002; Spring, 1997; Meyer & Schartl, 1999; Wang & Gu, 2000; Dehal & Boore, 2005), others find evidence of only one round of WGD (X.Gu et al., 2002; Guigo et al., 1996; McLysaght et al., 2002), while others disclaim the possibility of WGD entirely and discuss only duplications of a limited number of segments (Friedman & Hughes, 2001; 2003).

## 2.2 Smaller scale duplication

Ohno (1970) argued that duplication of the genome rather than its individual parts is more important for evolution, because partial duplications can lead to regulatory imbalances. Nevertheless, partial and complete duplications of genes also play very important roles in evolution. WGDs have occurred several times during the evolutionary history of organisms, while SSDs arise continuously through multiple mechanisms. Several mechanisms have been suggested for the improvement in function of existing proteins and for the creation of new functions. One such mechanism is the internal (partial) duplication of genes, which is important for increasing the functional complexity of genes in evolution (Li, 1997). Such duplications are believed to have played a key role in the emergence of complex genes. Many proteins of modern organisms contain internal repeats of amino acids, and these repeats often correspond to functional or structural domains of proteins. These data suggest that the genes encoding these proteins were formed by internal duplications (Lavorgna et al., 2001). Internal duplication provides the possibility of improving protein function by increasing the number of active sites. Internal duplications can also lead to the acquisition of new functions by the modification of duplicated regions or the reorganization of modules. Numerous data on the role of intragenic duplications in the early stages of evolution of proteins were obtained by comparative analyses of sequenced genomes (Marcotte et al., 1999; Lavorgna et al., 2001; Conant & Wagner 2005; Chen et al. 2007). Duplicated regions can accumulate mutations that contribute to the divergence of the repeated fragments, which can then become fixed. Often, only traces of duplications in the form of imperfect repeats can be detected in contemporary amino acid sequences (Li, 1997). Eukaryotic proteins have more repeats than do prokaryotic proteins (Marcotte et al., 1999; Chen et al., 2007).

## 3. The fate of duplicated genes

Tens of millions of years after WGD in *Arabidopsis thaliana* and *S. cerevisiae*, only about 30% and 10%, respectively, of the genes are preserved in the form of duplicated copies (Seoighe & Wolfe, 1999; Wong et al., 2002; Blanc et al., 2003). Preservation of duplicated copies in evolution can be achieved by one of three processes: (1) conservation, in which the copies are stored in an unaltered state (Hahn 2009); (2) subfunctionalization, in which both paralogs are necessary for performing the functions previously provided by the ancestral gene (both terms were offered by Force (Force et al., 1999)); and (3) neofunctionalization, in which one of the paralogs acquires a new function and the other preserves the old function. Characteristically, in (2) and (3), the regulatory and/or structural parts of the gene may be changed (Figure 1).

### 3.1 Conservation of duplicated copies

Duplicated genes are retained unchanged in cases where the normal development of the organism needs many copies of genes with similar function, which allows the synthesis of a

larger amount of specific RNA or protein (Ohno, 1970). An increase in the number of copies of these genes correlates with the increasing complexity of the organism (Chen et al., 2007). Amplification of genes in microorganisms leads to resistance to antibiotics and heavy metals, increased virulence and other adaptive properties (Romero & Palacios, 1997; Reams & Neidle, 2004; Andersson & Hughes, 2009). In plants, amplification of genes provides resistance to herbicides (Harms et al., 1992; Shyr et al., 1992). The best known examples of conservation of duplicated copies in various organisms are genes for rRNA, tRNA and histones, many of which are organized in tandem repeats, which allows the maintenance of homogeneity by unequal crossing over or gene conversion (Hurles, 2004).

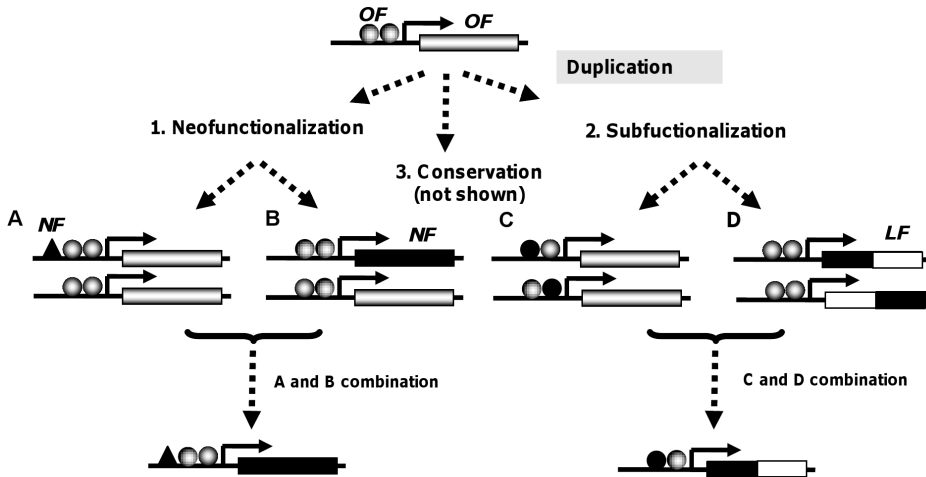


Fig. 1. Possible consequences of gene duplication (modified from (Hahn, 2009)). A and C - regulatory sequence changes; B and D - coding sequence changes. Since variant 3 (conservation) does not change the duplicated copies, it is not represented in the diagram. OF (grey) - old function, NF (black) - new function, LF (white) - lost function (attributed to both regulatory and structural sequences).

One of the most interesting questions related to the preservation of duplicated copies of genes is whether the loss of genes is an occasional event or is subjected to natural selection. Which duplicates are lost, and which persist after polyploidization? About 10% of yeast genes are preserved in the form of duplicated copies, and most are not needed for viability (Z.Gu et al., 2002). The most frequently duplicated genes encode cyclins, components of the signal transduction pathway, and cytoplasmic (but not mitochondrial) ribosomal proteins. Most are characterized by high levels of expression. Perhaps selection for increasing the level of expression was the major factor for the preservation of duplicated genes (Seoighe & Wolfe, 1999).

Analysis of the most recent WGD in *Arabidopsis* showed a preferential retention of genes involved in transcription and signal transduction, whereas genes involved in DNA repair or encoding proteins of organelles were characterized by more frequent loss (Blanc & Wolfe, 2004). Interestingly, genes preserved as paralogs after duplication have a high probability of remaining duplicated after the next round of duplication (Seoighe & Gehring, 2004). Loss of duplicates is thus not a random process.

### 3.2 Subfunctionalization

This hypothesis is to some extent the opposite to Ohno's hypothesis of evolution, because it assumes the existence of both functions before the duplication (Figure 1). The first evidence for it was the discovery of the phenomenon of "gene sharing" (Piatigorsky et al., 1988; Piatigorsky, 2003).

This model explains the emergence of new genes by the duplication of multifunctional genes. Such genes encode proteins that already perform different functions. Gene sharing was discovered in crystallins, proteins found in the lens of the eye. Crystallins make up 70% of the contents of cells, but remain in soluble form without forming aggregates (formation of aggregates leads to cataracts). Under such conditions, a majority of proteins would form insoluble aggregates within seconds. Another feature of these proteins is a record longevity (equal to the lifetime of an individual, for example 80 years); most proteins last for only minutes or hours. The eyes of all vertebrates have a standard set of crystallins ( $\alpha$ ,  $\beta$ ,  $\gamma$ ), and additional species-specific crystallins are encoded by genes that in other tissues encode enzymes. In most cases, this double life is ensured not by duplications but by a "division of functions": enzyme and crystallin are encoded by the same gene, but the protein can perform additional functions without changing its amino acid sequence. This phenomenon was thus called gene sharing (Piatigorsky et al., 1988; Piatigorsky, 2003). In gene sharing, a gene acquires a second function, without duplication and without loss of its primary function. A change in tissue specificity or regulation during development, however, may occur. Acquisition of a new function without duplication was first detected in crystallin  $\epsilon$  in birds and crocodiles (up to 23% of the total protein of the lens). The amino acid sequence of crystallin  $\epsilon$  was identical to lactate dehydrogenase B (LDH), and the protein had an activity similar to LDH. Subsequent work showed that both proteins were encoded by the same gene. Similarly, crystallin  $\tau$  in lampreys, bony fishes, reptiles and birds is identical to and encoded by the same gene as  $\alpha$  enolase. Zeta-crystallin is identical to quinone reductase. Crystallins  $\delta$ ,  $\epsilon$  and  $\tau$  thus illustrate examples of "division of functions", when a gene has acquired additional functions, without duplication. Multifunctional genes are characterized by significant limitations in the capabilities of any adaptive changes, since mutation that improves one function may disturb another. Duplication could provide a possible resolution of this "adaptive conflict". The molecular mechanisms leading to subfunctionalization have not been studied in detail until recently. Such analyses only became possible with the comparative analysis of genes in closely related species, for example in genes involved in galactose utilization in *S. cerevisiae* and *K. lactis* (Hittinger & Carroll, 2007). Divergence in the expression of duplicated genes over long periods of time attracted the interest of scientists as an important stage in the emergence of a new gene by duplication (Ohno, 1970; Ferris & Whitt, 1979). Thus in some cases, duplicates may have identical coding sequences but different regulatory sequences (Figure 1). Some pairs of duplicated genes can diverge in concert, forming two groups that are expressed in different tissues or under different conditions (Blanc & Wolfe, 2004). This process, which explains the divergence of metabolic pathways, is called "concerted divergence".

### 3.3 Neofunctionalization

The stable maintenance of duplicated copies in the genome requires functional divergence. From Ohno's (1970) position, functional divergence is achieved by ensuring that one copy of the gene retains the old function, while other copies acquire new

functions. An inevitable intermediate stage in this process would be the emergence of a pseudogene, as most mutations will disrupt or inactivate a gene rather than giving rise to new functions. Because this event is considered extremely unlikely, an extended hypothesis of neofunctionalization (NF) has been proposed, which includes the following possibilities: (1) a new gene acquires a new function but keeps the old function (NF-I), (2) a new gene completely loses the old function (NF-II), or (3) a new gene retains part of the old function (NF-III) (He & Zhang, 2005). Many examples of neofunctionalization have been described in recent years (see (Hahn, 2009)), although distinguishing neofunctionalization from subfunctionalization is sometimes difficult and has led to the creation of a "subneofunctionalization" model (He & Zhang, 2005).

### 3.4 Exon shuffling as a mechanism of neofunctionalization

One of the options for neofunctionalization is the formation of "chimeric" or fusion genes (Long, 2000). This phenomenon is possible due to the duplication of a gene or part of a gene, because only then can the original gene remain functional. After gene duplication, one of the copies can capture an exon(s) from an unrelated adjacent gene. Another possibility is the addition of flanking non-coding DNA as an additional open reading frame. The model, known as "exon shuffling" (Gilbert, 1978), suggests that recombination in introns can provide a mechanism for exchanging exon sequences between genes. However, the event will be evolutionarily significant only if it involves a structural or functional domain. Moreover, the shuffling of domains can occur without the involvement of introns (Doolittle, 1995). We are thus more correct to discuss the shuffling of domains rather than exons. Introns do not occur in prokaryotic genes, but many cases of domain shuffling have been described. The presence of introns, though, greatly facilitates the shuffling of domains, especially in vertebrates. In the 30 years since the discovery of introns, many examples of exon shuffling in a variety of organisms (vertebrates, invertebrates, plants) have been found. Only relatively recently have retrotransposition and illegal recombination been shown to be responsible for these phenomena (Long et al., 2003; van Rijk & Bloemendal, 2003).

## 4. Translation factors as examples of subneofunctionalization

### 4.1 The main stages of translation

In the process of protein synthesis, or translation, four distinct phases are usually distinguished: initiation, elongation, termination and recycling (Figure 2).

During **initiation**, the ribosome is assembled at the initiation codon of the mRNA, and the initiating methionyl-tRNA is attached to the peptidyl (P) center of the ribosome. The main objectives of the initiation of translation are identical in bacteria and eukaryotes, but initiation is much more complex in eukaryotes than in bacteria (Kapp & Lorsch, 2004). Three initiation factors occur in bacteria, but eukaryotes have at least 12, which contain about 23 different proteins (Sonenberg & Dever, 2003). Interestingly, the initiation of translation in archaea is intermediate in complexity between bacterial and eukaryotic translation.

During **elongation**, the aminoacyl-tRNA binds to the aminoacyl center (or A-site) of the ribosome, where the information recorded on the mRNA is translated into the language of proteins. This process involves elongation factor eEF1A (EF-Tu in bacteria) in complex with GTP. The ribosomes catalyze the formation of peptide bonds when the anticodons of tRNAs correspond to the codons of the mRNA. After translocation of the mRNA in the P-center, with the help of eEF2 (EF-G in bacteria), a next codon arrives in the A-center, and the

process repeats. In contrast to initiation, the main components involved in elongation are highly conserved in all three domains. For example, the human elongation factor eEF1A and EF-Tu of *Escherichia coli* are 33% identical along their entire length, exhibiting a higher degree of similarity in the GTP-binding domains (Cavallius *et al.* 1993). The proteins a/eEF1A and a/eEF2 reveal significant structural similarities, both in the free state and in complex with the ribosome (Andersen *et al.*, 2001; Stark *et al.*, 2002; Valle *et al.*, 2002; Jorgensen *et al.*, 2003). The similarity of elongation factors in bacteria, archaea and eukaryotes suggests that the mechanisms of elongation in eukaryotes in many respects correspond to those in bacteria and archaea (Ramakrishnan, 2002).

**Termination** of translation begins when the stop codon (UAA, UAG or UGA) enters the A-site of the ribosome. As the result of this process, the newly synthesized polypeptide chain is released. The stop codon is recognized by a release factor (RF1/RF2 in prokaryotes and eRF1 in eukaryotes) that triggers release of the nascent peptide from the ribosome. The efficiency of termination is enhanced by the GTPase release factor, RF3 in prokaryotes and eRF3 in eukaryotes (Kisselev *et al.*, 2003). At least some stages of the termination of translation, such as recognition of the stop codon and hydrolysis of peptidyl-tRNAs, are assumed to be similar in archaea and eukaryotes. This hypothesis is based on data of the homology of aRF1 and eRF1 and the finding that aRF1 of *Methanococcus jannaschii* is able to function in an *in vitro* system containing mammalian ribosomes (Dontsova *et al.*, 2000). Archaea, however, do not have homologs of RF3 and eRF3, which does not necessarily mean the absence of proteins with similar functions. Alternatively, these proteins may be absent due to a reduction of the apparatus of translation during the evolution of archaea (Lecompte *et al.*, 2002).

During the final stage of translation, **recycling**, the dissociation of the ribosome occurs together with the release of the mRNA and deacylated tRNAs. An essential feature of this stage is the preparation of a new round of initiation. The details of this process are known only for bacteria.

## 4.2 Termination factors have arisen by the duplication of genes encoding elongation factors

Comparison of amino acid sequences in the family of elongation factors raised speculation that the progenitors of EF-G and EF-Tu arose as a result of duplication and subsequent divergence of a gene encoding an ancient GTPase, and further duplications led to the emergence of modern elongation and termination factors (Nakamura & Ito, 1998; Inagaki & Doolittle, 2000) (Figure 3). RF1, RF2 and RF3, as well as eRF1 and elongation factor eEF-2, are assumed to have been derived from the bacterial elongation factor EF-G (Nakamura & Ito 1998), while eRF3 arose from the duplication of the gene encoding eukaryotic elongation factor eEF1-A (Inagaki & Doolittle 2000).

The amino acid sequences of RF1 and RF2 are 36% identical, suggesting that the genes *prfA* and *prfB* arose from a common precursor by duplication (Craigien *et al.*, 1990). Homologs of eRF1 are found in different species, and the eRF1 protein from different species is able to replace eRF1 of *S. cerevisiae*, indicating a high degree of functional conservation (Urbero *et al.*, 1997). An almost complete lack of similarity in the sequences of bacterial and eukaryotic termination factors probably indicates their independent origin (Kisselev *et al.*, 2003). On the other hand, the first class factors (RF1, RF2, aRF1 and eRF1) could be so divergent that they have lost any resemblance, with the exception of the GGQ motif (Frolova *et al.*, 1999; Lecompte *et al.*, 2002; Seit-Nebi *et al.*, 2001). The lack of homology between the amino acid

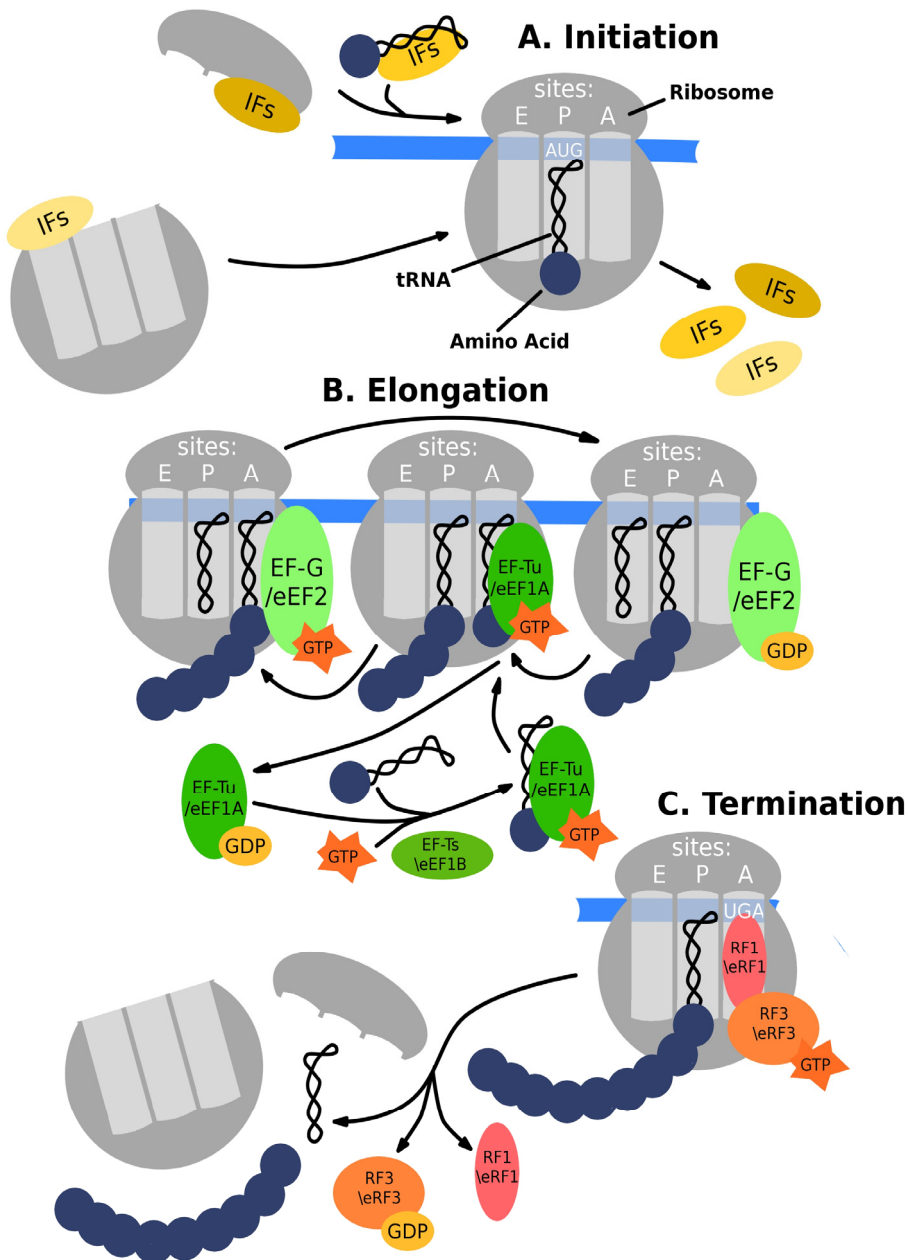


Fig. 2. Evolutionarily related proteins perform similar functions and interact with the same sites of the ribosome during translation. The most significant participants are shown. The arrows indicate the sequence of events. IF - initiation factor; EF - elongation factor; RF - release, or termination, factor; e - eukaryotic.

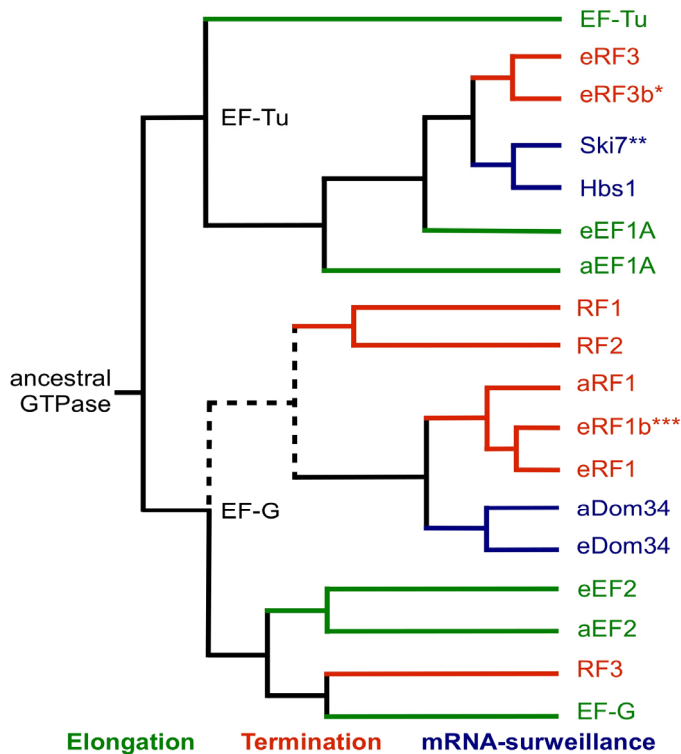


Fig. 3. The origin of the proteins involved in elongation, termination and mRNA quality control. The genes duplicated only in certain taxa are marked with asterisks: \* - duplication unique to mammals (Hoshino et al., 1998; Jakobsen et al., 2001); \*\* - duplication described only from *Saccharomyces* (Atkinson et al., 2008); \*\*\* - duplication specific to several species of ciliates (Liang et al., 2001; Atkinson et al., 2008) and *A. thaliana* (Chapman & Brown, 2004). Branch lengths are not to scale. The progenitors of prokaryotic EF-G and EF-Tu were proposed to have first diverged from a common ancestral GTPase, and then each gave rise to two protein families corresponding to the elongation and termination factors (Nakamura & Ito, 1998; Inagaki & Doolittle, 2000; Atkinson et al., 2008). EF - elongation factor, RF - release factor, e - eukaryotic, a - archaeal.

sequences of bacterial and eukaryotic termination factors does not mean that these proteins lack similarity at other levels of the organization of protein molecules. Indeed, the spatial structure of many translation factors are characterized by a number of common features that fit the hypothesis of "molecular mimicry" (Nissen et al., 2000; Nakamura & Ito, 2003).

In contrast to eRF1, eRF3 is a much less conserved protein, especially in its N-terminal domain, which can either be completely absent, as in the case of *Giardia lamblia* (Inagaki & Doolittle, 2000), or demonstrate species-specific differences in length (maximum length is 321 amino acids in *Leishmania major* (Atkinson et al., 2008)) and amino acid sequence. This lack of conservation may underlie species-specific regulation of the activity of this protein (Kodama et al., 2007). In some species of yeast, the N-terminus is enriched in QN residues

and provides prionogenic properties to the protein (Kushnirov & Ter Avanesyan, 1998). The same amino acid composition is also detected in the N-terminal domains of eRF3 in the kinetoplastid protists *L. major* and *Trypanosoma cruzi*, but this similarity is unlikely to be homologous (Atkinson et al., 2008). For termination of translation and maintenance of viability, only the C-terminal domain of eRF3 (homologous to elongation factor eEF1A) is necessary. eRF3 may have arisen in the early stages of eukaryotic evolution, since neither bacterial nor archaeal genomes contain homologues of eRF3 (Inagaki & Doolittle, 2000). Recent studies have shown that the functions of eRF3 can be performed in archaea by aEF1A (Saito et al., 2010).

The termination factor eRF3, preserving the functions typical of elongation factors (GTP-ase activity and interaction with the A-site of the ribosome), lost the capacity to bind tRNA but acquired the capacity to interact with eRF1 (Table 1). From this standpoint, elongation factor EF1A of archaea is functionally intermediate between elongation and termination factors: it acquired the ability to stimulate aRF1 while maintaining all the properties of an elongation factor (Saito et al., 2010). Termination factor eRF1 is a striking example of neofunctionalization, because it has acquired a variety of functions absent in elongation factors, including the ability to decode stop signals and to catalyze the release of nascent peptides from eukaryotic ribosomes in response to stop codons.

	GTP-binding	tRNA binding and delivering to the A-site of ribosome	Recognition of stop-signal in A-site of ribosome	Function of accessory protein
<b>Archaea</b>	aEF1A	aEF1A	aRF1	aEF1A (for aRF1), aEF1B (for aEF1A)
<b>Eubacteria</b>	EF-Tu, RF3, EF-G	EF-Tu	RF1, RF2	RF3 (for RF1 or RF2), EF-Ts (for EF-Tu)
<b>Eucarya</b>	eEF1A, eRF3, eEF-2	eEF1A	eRF1	eRF3 (for eRF1), eEF1B (for eEF1A)

Table 1. Functional homology between elongation and termination factors in Archaea, Bacteria and Eukaryota

### 4.3 Additional paralogs of termination factors in several species

Additional duplication of genes encoding termination factors have been found in several species (Figure 3). For example, an additional copy of eRF1 is present in some lineages of ciliates (Liang et al., 2001; Atkinson et al., 2008). These organisms differ from most eukaryotes by their reassignment of one or two stop codons to encode amino acids (Lozupone et al., 2001). UGA, for instance, encodes cysteine in *Euplotes* (Meyer et al., 1991). The presence of two copies of eRF1 in *Euplotes octocarinatus* may be associated with a different codon specificity of eRF1 proteins for UAA and UAG codons (Liang et al., 2001). Later studies showed that both eRF1a and eRF1b recognized UAA and UAG as stop codons



(Wang et al., 2010). The precise functions of each protein thus remain to be discovered. The plant *A. thaliana* has three paralogs of eRF1, all of which are able to rescue the *sup45-2(ts)* mutation in *SUP45* (encoding eRF1) in *S. cerevisiae* (Chapman & Brown, 2004).

Another example of duplication, found only in some taxonomic groups, is the presence of two paralogous genes encoding eRF3 in mammals. In mammals, proteins homologous to eRF3 can be divided into two subfamilies based on the sequence of their N-termini. The first subfamily includes human hGSPT1 (or eRF3a) and mouse mGSPT1 (Hoshino et al., 1989; Hoshino et al., 1998; Jean-Jean et al., 1996), while the second subfamily includes human hGSPT2 (eRF3b) and mouse mGSPT2 (Hoshino et al., 1998; Jakobsen et al., 2001). Complementation experiments have shown that only *mGSPT2* is able to complement the *SUP35* gene (encoding eRF3) mutation (Le Goff et al., 2002). *GSPT2* is a paralog of *GSPT1* that has perhaps arisen as a result of retrotransposition of the *GSPT1* transcript into the genome of the common ancestor of mouse and human. *GSPT2* may thus be a functional retrogene (Zhouravleva et al., 2006). Both eRF3a and eRF3b are able to serve as termination factors in mammalian cells and interact with eRF1 (Chauvin et al., 2005). However, eRF3a is considered the main factor (Chauvin et al., 2005) that is expressed in all tissues, while eRF3b is detected only in the brain (Hoshino et al., 1998; Chauvin et al., 2005). This duplication event may not have led to the emergence of a new gene function but may have contributed to the complexity of regulatory processes by tissue-specific expression of these genes.

#### **4.4 Subneofunctionalization in a family of termination factors gave rise to proteins participating in mRNA quality control**

A necessary condition of protein synthesis is to obtain functionally active proteins, so the control of accuracy of protein synthesis occurs at each stage of translation (Valente & Kinzy, 2003). The accuracy of initiation is achieved by proper identification of the start codon by a multifactorial initiation complex (Asano et al., 2001). Elongation requires the control of various events, including maintenance of the correct reading frame. Shifts in the reading frame occur at a frequency near  $3 \times 10^{-5}$  (Atkins et al., 1991) and may lead to the synthesis of non-functional products because shifts in the reading frame will often create a premature termination codon (PTC).

Eukaryotic cells possess a mechanism known as nonsense-mediated mRNA decay (**NMD**) that recognizes and degrades mRNA molecules containing premature termination codons (Amrani et al., 2006) (Figure 4). NMD is mediated by the trans-acting factors Upf1, Upf2 and Upf3, all of which directly interact with eRF3; only Upf1 interacts with eRF1 (Czaplinski et al., 1998; Wang et al., 2001). In addition to NMD, eukaryotic cells contain two additional mechanisms of mRNA quality control. No-go decay (**NGD**) releases ribosomes that are stalled on the mRNA (Doma & Parker, 2006). In yeast, NGD involves the proteins Hbs1 and Dom34 (Pelota in mammals). Another mechanism, non-stop decay (**NSD**), leads to the release of ribosomes that have read through the stop codon instead of terminating (Vasudevan et al., 2002). NSD has only been found in *S. cerevisiae* and involves the Ski7 protein (van Hoof et al., 2002). A common feature of these processes is that all involve the termination factors eRF1 and eRF3 (NMD) or their paralogs (Dom34/eRF1 and Hbs1/eRF3 in NGD; Ski7/eRF3 in NSD).

Hbs1 is a paralog of eEF1A and eRF3 (Wallrapp et al., 1998; Inagaki & Doolittle, 2000), while Dom34 is a paralog of eRF1 (Koonin et al., 1994; Davis & Engebrecht, 1998) (Figure 3). The C-terminus of Hbs1, homologous to that of eRF3, is sufficient to interact with Dom34, which assumes the same structure of the complex of two pairs of proteins (Hbs1-Dom34 and eRF3-

eRF1) (Carr-Schmid et al., 2002). Indeed, Hbs1 forms a complex with Dom34 and GTP (Dom34-Hbs1-GTP), similar to that of eRF1-eRF3-GTP (Hauryliuk et al., 2006; Graille et al., 2008; Chen et al., 2010; Shoemaker et al., 2010; van den Elzen et al., 2010). The central event of NGD is mRNA cleavage, and Dom34 has the necessary RNase activity (Lee et al., 2007; Graille et al., 2008), although the proposed endonuclease activity of Dom34 is not required for mRNA cleavage in NGD (Passos et al., 2009). Dom34 of *S. cerevisiae* consists of three domains, two of which are homologous to the corresponding domains in eRF1, while the N-terminal domain of Dom34 is different from that of eRF1 and is probably necessary for the

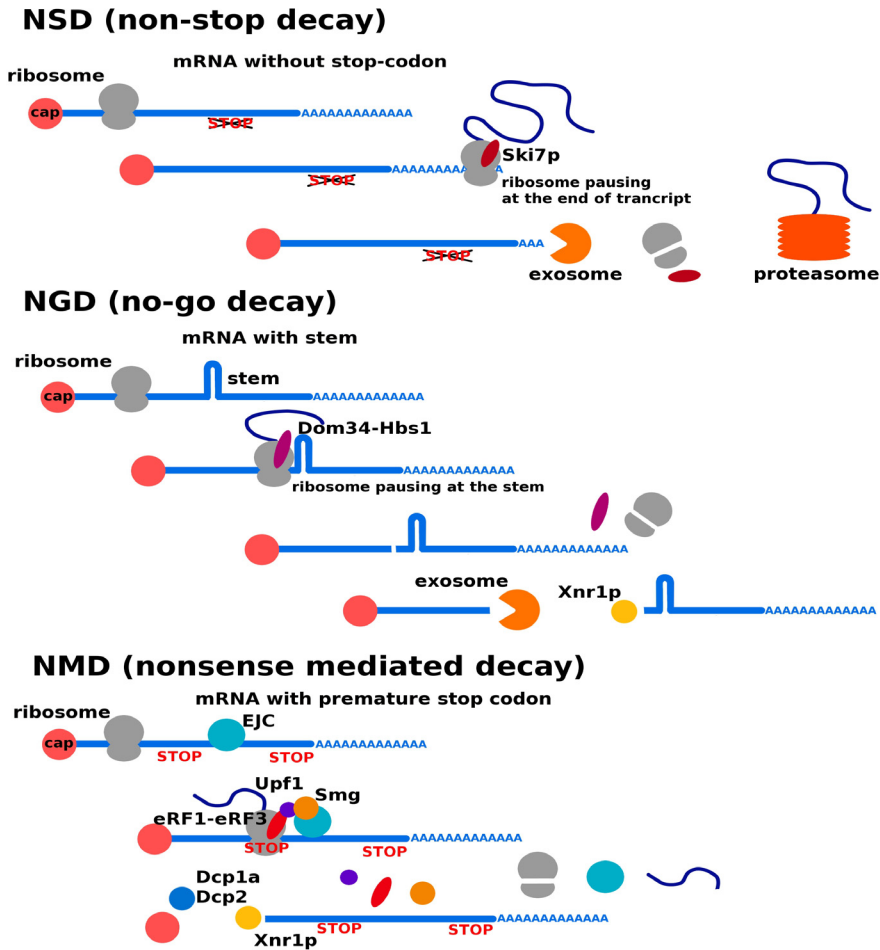


Fig. 4. Neofunctionalization of termination factors in mRNAs quality control systems. Three systems described for *S. cerevisiae* are shown. NSD (Non-stop decay) is responsible for the degradation of transcripts lacking stop codons. NGD (No-go decay) removes mRNA secondary structures that prevent translation. NMD (Nonsense-mediated decay) destroys transcripts containing nonsense mutations. See text for details.

recognition of the mRNA stem (Graille et al., 2008). Lack of the Hbs1 protein in archaea is apparently compensated by its homolog aEF1A (Kobayashi et al., 2010), which also performs the functions of eRF3 in archaeal termination of translation (Saito et al., 2010). In one more pathway of mRNA degradation, non-stop decay (NSD), participates In one more pathway of mRNA degradation, non-stop decay (NSD), participates Ski7 protein that is paralog of Hbs1 and eRF3 (Benard et al., 1999). This mechanism is necessary to destroy mRNAs lacking all termination codons (Frischmeyer et al., 2002; van Hoof et al., 2002). Ski7 protein that is paralog of Hbs1 and eRF3 (Benard et al., 1999). This mechanism is necessary to destroy mRNAs lacking all termination codons (Frischmeyer et al., 2002; van Hoof et al., 2002) Ski7, involved in NSD, arose from duplication of Hbs1 by WGD (Kellis et al., 2004) or by an independent duplication of Hbs1 before WGD and the subsequent loss in several species (Atkinson et al., 2008) (Figure 3). An interesting hypothesis links the appearance of Ski7 with the existence of the prion  $[PSI^+]$  (Atkinson et al., 2008).  $[PSI^+]$  is the aggregated (prion) form of the yeast protein Sup35 (eRF3) (Kushnirov & Ter Avanesyan, 1998). Formation of  $[PSI^+]$  decreases the amount of functional Sup35, leading to the efficient read-through of nonsense mutations in ORFs (and possibly at the normal terminator codons) (Serio & Lindquist, 1999). The emergence of Ski7 in such organisms would thus create an additional system of mRNA quality control. However,  $[PSI^+]$  formation has not been detected in the natural, industrial and clinical isolates of *Saccharomyces*. In addition, the prionic properties of Sup35 are conserved in various species of *Saccharomyces* as well as in *Candida albicans* and *Pichia methanolica* (Inge-Vechtomov et al., 2003), species in which Ski7 has not been found (Atkinson et al., 2008).

## 5. Conclusion

Successive duplications of genes encoding elongation factors for translation led to the emergence of several protein complexes with different properties. The eRF1-eRF3 complex terminates translation, and the Dom34-Hbs1 complex is involved in the quality control of mRNA. Both eRF1 and eRF3 interact not only with each other but also with additional proteins. Some of these interactions are possibly mutually exclusive, and some of the proteins interacting with eRF1/eRF3 can be components of the complex terminating translation. Possible candidates for involvement in termination are poly(A) binding protein (PABP) and Upf proteins (Upf1, Upf2 and Upf3). Interaction of eRF3 with PABP links termination of translation with initiation (Hoshino et al., 1999), while interaction with Upf involves eRF proteins in nonsense-mediated decay (Amrani et al., 2006). The genetic data, derived mostly from *S. cerevisiae*, strongly suggest that the functions of eRF1 and eRF3 are not restricted to termination of translation (Inge-Vechtomov et al., 2003). Further studies are needed to characterize other non-translational functions of both proteins, as was shown for eEF1A (Mateyak & Kinzy, 2010).

## 6. Acknowledgments

This work was supported by the Russian Foundation for Basic Research (10-04-00237) and the Program of the Presidium of the Russian Academy of Sciences, The Origin and the Evolution of the Biosphere.

## 7. References

- Adams, K.L. & Wendel, J.F. (2005) Polyploidy and genome evolution in plants. *Curr Opin.Plant Biol.*, Vol. 8, pp. 135-141.
- Amrani, N., Dong, S., He, F., Ganesan, R., Ghosh, S., Kervestin, S., Li, C., Mangus, D.A., Spatrnick, P., & Jacobson, A. (2006) Aberrant termination triggers nonsense-mediated mRNA decay. *Biochem.Soc.Trans.*, Vol. 34, pp. 39-42.
- Andersen, G.R., Valente, L., Pedersen, L., Kinzy, T.G., & Nyborg, J. (2001) Crystal structures of nucleotide exchange intermediates in the eEF1A-eEF1B $\alpha$  complex. *Nat.Struct.Biol.*, Vol.8, pp. 531-534.
- Andersson, D.I. & Hughes, D. (2009) Gene Amplification and Adaptive Evolution in Bacteria. *Annu.Rev.Genet.*, Vol.43, pp. 167-95.
- Asano, K., Phan, L., Valasek, L., Schoenfeld, L.W., Shalev, A., Clayton, J., Nielsen, K., Donahue, T.F., & Hinnebusch, A.G. (2001) A multifactor complex of eIF1, eIF2, eIF3, eIF5, and tRNA(i)Met promotes initiation complex assembly and couples GTP hydrolysis to AUG recognition. *Cold Spring Harb.Symp.Quant.Biol.*, Vol.66, pp. 403-415.
- Atkins, J.F., Weiss, R.B., Thompson, S., & Gesteland, R.F. (1991) Towards a genetic dissection of the basis of triplet decoding, and its natural subversion: programmed reading frame shifts and hops. *Annu.Rev.Genet.*, Vol.25, pp. 201-228.
- Atkinson, G.C., Baldauf, S.L., & Hauryliuk, V. (2008) Evolution of nonstop, no-go and nonsense-mediated mRNA decay and their termination factor-derived components. *BMC.Evol.Biol.*, Vol.8, pp. 290.
- Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Segurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Camara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A.M., Kissmehl, R., Klotz, C., Koll, F., Le Mouel, A., Lepere, G., Malinsky, S., Nowacki, M., Nowak, J.K., Plattner, H., Poulain, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Betermier, M., Weissenbach, J., Scarpelli, C., Schachter, V., Sperling, L., Meyer, E., Cohen, J., & Wincker, P. (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*, Vol.444, pp. 171-178.
- Benard, L., Carroll, K., Valle, R.C., Masison, D.C., & Wickner, R.B. (1999) The Ski7 antiviral protein is an EF1- $\alpha$  homolog that blocks expression of non-Poly(A) mRNA in *Saccharomyces cerevisiae*. *J.Virol.*, Vol.73, pp. 2893-2900.
- Blanc, G., Hokamp, K., & Wolfe, K.H. (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.*, Vol.13, pp. 137-144.
- Blanc, G. & Wolfe, K.H. (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, Vol.16, pp. 1667-1678.
- Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., & Van de, P.Y. (2006) The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.*, Vol.7, R43.
- Brunet, F.G., Crollius, H.R., Paris, M., Aury, J.M., Gibert, P., Jaillon, O., Laudet, V., & Robinson-Rechavi, M. (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol.Biol.Evol.*, Vol.23, pp. 1808-1816.

- Byrne, K.P. & Wolfe, K.H. (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.*, Vol.15, pp. 1456-1461.
- Carr-Schmid, A., Pfund, C., Craig, E.A., & Kinzy, T.G. (2002) Novel G-protein complex whose requirement is linked to the translational status of the cell. *Mol.Cell Biol.*, Vol.22, pp. 2564-2574.
- Cavallius, J., Zoll, W., Chakraborty, K., & Merrick, W.C. (1993) Characterization of yeast EF-1  $\alpha$ : non-conservation of post-translational modifications. *Biochim.Biophys.Acta*, Vol.1163, pp. 75-80.
- Chapman, B. & Brown, C. (2004) Translation termination in *Arabidopsis thaliana*: characterisation of three versions of release factor 1. *Gene*, Vol.341, pp. 219-225.
- Chapman, B.A., Bowers, J.E., Feltus, F.A., & Paterson, A.H. (2006) Buffering of crucial functions by paleologous duplicated genes may contribute cyclicity to angiosperm genome duplication. *Proc.Natl.Acad.Sci.U.S.A.*, Vol.103, pp.2730-2735.
- Chauvin, C., Salhi, S., Le Goff, C., Viranaicken, W., Diop, D., & Jean-Jean, O. (2005) Involvement of human release factors eRF3a and eRF3b in translation termination and regulation of the termination complex formation. *Mol.Cell Biol.*, Vol.25, pp. 5801-5811.
- Chen, C.C., Li, W.H., & Sung, H.M. (2007) Patterns of internal gene duplication in the course of metazoan evolution. *Gene*, Vol.396, pp. 59-65.
- Chen, L., Muhrlad, D., Hauryliuk, V., Cheng, Z., Lim, M.K., Shyp, V., Parker, R., & Song, H. (2010) Structure of the Dom34-Hbs1 complex and implications for no-go decay. *Nat.Struct.Mol.Biol.* 17, 1233-1240.
- Comai, L., Tyagi, A.P., Winter, K., Holmes-Davis, R., Reynolds, S.H., Stevens, Y., & Byers, B. (2000) Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids. *Plant Cell*, Vol.12, pp. 1551-1568.
- Conant, G.C. & Wagner, A. (2005) The rarity of gene shuffling in conserved genes. *Genome Biol.*, Vol.6, R50.
- Craigen, W.J., Lee, C.C., & Caskey, C.T. (1990) Recent advances in peptide chain termination. *Mol.Microbiol.*, Vol.4, pp. 861-865.
- Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., Albert, V.A., Ma, H., & dePamphilis, C.W. (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Res.*, Vol.16, pp. 738-749.
- Czaplinski, K., Ruiz-Echevarria, M.J., Paushkin, S.V., Han, X., Weng, Y., Perlick, H.A., Dietz, H.C., Ter Avanesyan, M.D., & Peltz, S.W. (1998) The surveillance complex interacts with the translation release factors to enhance termination and degrade aberrant mRNAs. *Genes Dev.*, Vol.12, pp. 1665-1677.
- Davis, J.C. & Petrov, D.A. (2005) Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet.*, Vol.21, pp. 548-551.
- Davis, L. & Engebrecht, J. (1998) Yeast *dom34* mutants are defective in multiple developmental pathways and exhibit decreased levels of polyribosomes. *Genetics*, Vol.149, pp. 45-56.

- Dehal, P. & Boore, J.L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.*, Vol.3, e314.
- Dietrich, F.S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pohlmann, R., Luedi, P., Choi, S., Wing, R.A., Flavier, A., Gaffney, T.D., & Philippsen, P. (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science*, Vol.304, pp. 304-307.
- Doma, M.K. & Parker, R. (2006) Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. *Nature*, Vol.440, pp. 561-564.
- Dontsova, M., Frolova, L., Vassilieva, J., Piendl, W., Kisselev, L., & Garber, M. (2000) Translation termination factor aRF1 from the archaeon *Methanococcus jannaschii* is active with eukaryotic ribosomes. *FEBS Lett.*, Vol.472, pp. 213-216.
- Doolittle, R.F. (1995) The multiplicity of domains in proteins. *Annu.Rev.Biochem.*, Vol.64, pp. 287-314.
- Durand, D. & Hoberman, R. (2006) Diagnosing duplications-can it be done? *Trends Genet.*, Vol.22, pp. 156-164.
- Ferris, S.D. & Whitt, G.S. (1979) Evolution of the differential regulation of duplicate genes after polyploidization. *J.Mol.Evol.*, Vol.12, pp. 267-317.
- Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., & Postlethwait, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, Vol.151, pp. 1531-1545.
- Friedman, R. & Hughes, A.L. (2001) Pattern and timing of gene duplication in animal genomes. *Genome Res.*, Vol.11, pp. 1842-1847.
- Friedman, R. & Hughes, A.L. (2003) The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol.Biol.Evol.*, Vol.20, pp. 154-161.
- Frischmeyer, P.A., van Hoof, A., O'Donnell, K., Guerrierio, A.L., Parker, R., & Dietz, H.C. (2002) An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. *Science*, Vol.295, pp. 2258-2261.
- Frolova, L.Y., Tsivkovskii, R.Y., Sivolobova, G.F., Oparina, N.Y., Serpinsky, O.I., Blinov, V.M., Tatkov, S.I., & Kisselev, L.L. (1999) Mutations in the highly conserved GGQ motif of class 1 polypeptide release factors abolish ability of human eRF1 to trigger peptidyl-tRNA hydrolysis. *RNA*, Vol.5, pp. 1014-1020.
- Gilbert, W. (1978) Why genes in pieces? *Nature*, Vol.271, pp. 501.
- Graille, M., Chaillet, M., & van Tilbeurgh, H. (2008) Structure of yeast Dom34: a protein related to translation termination factor eRF1 and involved in No-Go decay. *J.Biol.Chem.*, Vol.283, No.11, pp. 7145-7153.
- Gu, X., Wang, Y., & Gu, J. (2002) Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat.Genet.*, Vol.31, pp. 205-209.
- Gu, Z., Cavalcanti, A., Chen, F.C., Bouman, P., & Li, W.H. (2002) Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol.Biol.Evol.*, Vol.19, pp. 256-262.
- Guigo, R., Muchnik, I., & Smith, T.F. (1996) Reconstruction of ancient molecular phylogeny. *Mol.Phylogenet.Evol.*, Vol.6, pp. 189-213.

- Hahn, M.W. (2009) Distinguishing among evolutionary models for the maintenance of gene duplicates. *J.Hered.* Vol.100, pp. 605-617.
- Haldane, J. (1932) *The causes of evolution*, NY: Longmans, Green.
- Harms, C.T., Armour, S.L., DiMaio, J.J., Middlesteadt, L.A., Murray, D., Negrotto, D.V., Thompson-Taylor, H., Weymann, K., Montoya, A.L., Shillito, R.D., et al. (1992) Herbicide resistance due to amplification of a mutant acetohydroxyacid synthase gene. *Mol.Gen.Genet.*, Vol.233, pp. 427-435.
- Hauryliuk, V., Zavialov, A., Kisselev, L., & Ehrenberg, M. (2006) Class-1 release factor eRF1 promotes GTP binding by class-2 release factor eRF3. *Biochimie.*, Vol.88, pp. 747-757.
- He, X. & Zhang, J. (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, Vol.169, pp. 1157-1164.
- Hittinger, C.T. & Carroll, S.B. (2007) Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, Vol.449, pp. 677-681.
- Hoshino, S., Miyazawa, H., Enomoto, T., Hanaoka, F., Kikuchi, Y., Kikuchi, A., & Ui, M. (1989) A human homologue of the yeast *GST1* gene codes for a GTP-binding protein and is expressed in a proliferation-dependent manner in mammalian cells. *EMBO J.*, Vol.8, pp. 3807-3814.
- Hoshino, S., Imai, M., Mizutani, M., Kikuchi, Y., Hanaoka, F., Ui, M., & Katada, T. (1998) Molecular cloning of a novel member of the eukaryotic polypeptide chain-releasing factors (eRF). Its identification as eRF3 interacting with eRF1. *J.Biol.Chem.*, Vol.273, pp. 22254-22259.
- Hoshino, S., Imai, M., Kobayashi, T., Uchida, N., & Katada, T. (1999) The eukaryotic polypeptide chain releasing factor (eRF3/GSPT) carrying the translation termination signal to the 3'-Poly(A) tail of mRNA. Direct association of eRF3/GSPT with polyadenylate-binding protein. *J.Biol.Chem.*, Vol.274, pp. 16677-16680.
- Hurles, M. (2004) Gene duplication: the genomic trade in spare parts. *PLoS Biol.*, Vol.2, E206.
- Inagaki, Y. & Doolittle, F.W. (2000) Evolution of the eukaryotic translation termination system: origins of release factors. *Mol.Biol.Evol.*, Vol.17, pp. 882-889.
- Inge-Vechtomov, S., Zhouravleva, G., & Philippe, M. (2003) Eukaryotic release factors (eRFs) history. *Biol.Cell*, Vol.95, pp. 195-209.
- Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyere, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gasparo, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le, C., I, Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pe, M.E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.F., Weissenbach, J., Quetier, F., & Wincker, P. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, Vol.449, pp. 463-467.

- Jakobsen, C.G., Seggaard, T.M., Jean-Jean, O., Frolova, L., & Justesen, J. (2001) Identification of a novel termination release factor eRF3b expressing the eRF3 activity in vitro and in vivo. *Mol.Biol.(Mosk)*, Vol.35, pp. 672-681.
- Jean-Jean, O., Le Goff, X., & Philippe, M. (1996) Is there a human [psi]? *C.R.Acad.Sci.III.*, Vol.319, pp. 487-492.
- Jorgensen, R., Ortiz, P.A., Carr-Schmid, A., Nissen, P., Kinzy, T.G., & Andersen, G.R. (2003) Two crystal structures demonstrate large conformational changes in the eukaryotic ribosomal translocase. *Nat.Struct.Biol.*, Vol. 10, pp. 379-385.
- Kapp, L.D. & Lorsch, J.R. (2004) The molecular mechanics of eukaryotic translation. *Annu.Rev.Biochem.*, Vol.73, pp. 657-704.
- Kashkush, K., Feldman, M., & Levy, A.A. (2002) Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics.*, Vol.160, pp. 1651-1659.
- Kellis, M., Birren, B.W., & Lander, E.S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, Vol.428, pp. 617-624.
- Kisselev, L., Ehrenberg, M., & Frolova, L. (2003) Termination of translation: interplay of mRNA, rRNAs and release factors? *EMBO J.*, Vol.22, pp. 175-182.
- Kobayashi, K., Kikuno, I., Kuroha, K., Saito, K., Ito, K., Ishitani, R., Inada, T., & Nureki, O. (2010) Structural basis for mRNA surveillance by archaeal Pelota and GTP-bound EF1alpha complex. *Proc.Natl.Acad.Sci.U.S.A.*, Vol.107, pp. 17575-17579.
- Kodama, H., Ito, K., & Nakamura, Y. (2007) The role of N-terminal domain of translational release factor eRF3 for the control of functionality and stability in *S. cerevisiae*. *Genes Cells*, Vol.12, pp. 639-650.
- Koonin, E.V., Bork, P., & Sander, C. (1994) A novel RNA-binding motif in omnipotent suppressors of translation termination, ribosomal proteins and a ribosome modification enzyme? *Nucleic Acids Res.*, Vol.22, pp. 2166-2167.
- Kushnirov, V.V. & Ter Avanesyan, M.D. (1998) Structure and replication of yeast prions. *Cell*, Vol. 94, pp. 13-16.
- Larhammar, D., Lundin, L.G., & Hallbook, F. (2002) The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Res.*, Vol.12, pp. 1910-1920.
- Lavorgna, G., Patthy, L., & Boncinelli, E. (2001) Were protein internal repeats formed by "bricolage"? *Trends Genet*, Vol.17, pp. 120-123.
- Le Goff, C., Zemlyanko, O., Moskalenko, S., Berkova, N., Inge-Vechtomov, S., Philippe, M., & Zhouravleva, G. (2002) Mouse GSPT2, but not GSPT1, can substitute for yeast eRF3 in vivo. *Genes Cells*, Vol.7, pp. 1043-1057.
- Lecompte, O., Ripp, R., Thierry, J.C., Moras, D., & Poch, O. (2002) Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res.*, Vol.30, pp. 5382-5390.
- Lee, H.H., Kim, Y.S., Kim, K.H., Heo, I., Kim, S.K., Kim, O., Kim, H.K., Yoon, J.Y., Kim, H.S., Kim, d.J., Lee, S.J., Yoon, H.J., Kim, S.J., Lee, B.G., Song, H.K., Kim, V.N., Park, C.M., & Suh, S.W. (2007) Structural and functional insights into Dom34, a key component of no-go mRNA decay. *Mol.Cell.*, Vol.27, pp. 938-950.



- Lee, H.S. & Chen, Z.J. (2001) Protein-coding genes are epigenetically regulated in *Arabidopsis* polyploids. *Proc.Natl.Acad.Sci.U.S.A.*, Vol.98, pp. 6753-6758.
- Lewis, E.B. (1951) Pseudoallelism and gene evolution. *Cold Spring Harb.Symp.Quant.Biol.*, Vol.16, pp. 159-174.
- Li, W.H. (1997) *Molecular evolution* Sunderland, Mass: Sinauer Associates.
- Liang, A., Brunen-Nieweler, C., Muramatsu, T., Kuchino, Y., Beier, H., & Heckmann, K. (2001) The ciliate *Euplotes octocarinatus* expresses two polypeptide release factors of the type eRF1. *Gene*, Vol.262, pp. 161-168.
- Long, M. (2000) A new function evolved from gene fusion. *Genome Res.*, Vol.10, pp. 1655-1657.
- Long, M., Betran, E., Thornton, K., & Wang, W. (2003) The origin of new genes: glimpses from the young and old. *Nat.Rev.Genet.*, Vol.4, pp. 865-875.
- Lozupone, C.A., Knight, R.D., & Landweber, L.F. (2001) The molecular basis of nuclear genetic code change in ciliates. *Curr.Biol.*, Vol.11, pp. 65-74.
- Lynch, M. & Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, Vol.290, pp. 1151-1155.
- Lynch, M., O'Hely, M., Walsh, B., & Force, A. (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics*, Vol.159, pp. 1789-1804.
- Marcotte, E.M., Pellegrini, M., Yeates, T.O., & Eisenberg, D. (1999) A census of protein repeats. *J.Mol.Biol.*, Vol.293, pp. 151-160.
- Mateyak, M.K. & Kinzy, T.G. (2010) eEF1A: thinking outside the ribosome. *J.Biol.Chem.*, Vol.285, pp. 21209-21213.
- McLysaght, A., Hokamp, K., & Wolfe, K.H. (2002) Extensive genomic duplication during early chordate evolution. *Nat.Genet.*, Vol.31, pp. 200-204.
- Meyer, A. & Schartl, M. (1999) Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr Opin.Cell Biol.*, Vol.11, pp. 699-704.
- Meyer, F., Schmidt, H.J., Plumper, E., Hasilik, A., Mersmann, G., Meyer, H.E., Engstrom, A., & Heckmann, K. (1991) UGA is translated as cysteine in pheromone 3 of *Euplotes octocarinatus*. *Proc.Natl.Acad.Sci.U.S.A.*, Vol.88, pp. 3758-3761.
- Muller, H.J. (1936) Bar duplication. *Science*, Vol.83, pp. 528-530.
- Nakamura, Y. & Ito, K. (1998) How protein reads the stop codon and terminates translation. *Genes Cells*, Vol.3, pp. 265-278.
- Nakamura, Y. & Ito, K. (2003) Making sense of mimic in translation termination. *Trends Biochem.Sci.*, Vol.28, pp. 99-105.
- Nissen, P., Kjeldgaard, M., & Nyborg, J. (2000) Macromolecular mimicry. *EMBO J.*, Vol.19, pp. 489-495.
- Ohno, S. (1970). *Evolution by Gene Duplication*. NY: Springer Verlag.
- Otto, S.P. & Whitton, J. (2000) Polyploid incidence and evolution. *Annu.Rev.Genet.*, Vol.34, pp. 401-437.
- Passos, D.O., Doma, M.K., Shoemaker, C.J., Muhlrad, D., Green, R., Weissman, J., Hollien, J., & Parker, R. (2009) Analysis of Dom34 and its function in no-go decay. *Mol.Biol.Cell.*, Vol.20, pp. 3025-3032.

- Piatigorsky, J. (2003) Crystallin genes: specialization by changes in gene regulation may precede gene duplication. *J.Struct.Funct.Genomics*, Vol.3, pp. 131-137.
- Piatigorsky, J., O'Brien, W.E., Norman, B.L., Kalumuck, K., Wistow, G.J., Borrás, T., Nickerson, J.M., & Wawrousek, E.F. (1988) Gene sharing by delta-crystallin and argininosuccinate lyase. *Proc.Natl.Acad.Sci.U.S.A.*, Vol.85, pp. 3479-3483.
- Piskur, J. (2001) Origin of the duplicated regions in the yeast genomes. *Trends Genet.*, Vol.17, pp. 302-303.
- Ramakrishnan, V. (2002) Ribosome structure and the mechanism of translation. *Cell*, Vol.108, pp. 557-572.
- Reams, A.B. & Neidle, E.L. (2004) Selection for gene clustering by tandem duplication. *Annu.Rev.Microbiol.*, Vol.58, pp. 119-142.
- Rodin, S.N. & Riggs, A.D. (2003) Epigenetic silencing may aid evolution by gene duplication. *J.Mol.Evol.*, Vol.56, pp. 718-729.
- Romero, D. & Palacios, R. (1997) Gene amplification and genomic plasticity in prokaryotes. *Annu.Rev.Genet.*, Vol.31, pp. 91-111.
- Saito, K., Kobayashi, K., Wada, M., Kikuno, I., Takusagawa, A., Mochizuki, M., Uchiumi, T., Ishitani, R., Nureki, O., & Ito, K. (2010) Omnipotent role of archaeal elongation factor 1 alpha (EF1alpha) in translational elongation and termination, and quality control of protein synthesis. *Proc.Natl.Acad.Sci.U.S.A.* Vol.107, pp. 19242-19247.
- Scannell, D.R., Butler, G., & Wolfe, K.H. (2007) Yeast genome evolution--the origin of the species. *Yeast*, Vol.24, pp. 929-942.
- Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S., & Wolfe, K.H. (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, Vol.440, pp. 341-345.
- Seit-Nebi, A., Frolova, L., Justesen, J., & Kisselev, L. (2001) Class-1 translation termination factors: invariant GGQ minidomain is essential for release activity and ribosome binding but not for stop codon recognition. *Nucleic Acids Res.*, Vol.29, pp. 3982-3987.
- Seoighe, C. & Gehring, C. (2004) Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends Genet.*, Vol.20, pp. 461-464.
- Seoighe, C. & Wolfe, K.H. (1999) Yeast genome evolution in the post-genome era. *Curr Opin.Microbiol.*, Vol.2, pp. 548-554.
- Serio, T.R. & Lindquist, S.L. (1999) [PSI<sup>+</sup>]: an epigenetic modulator of translation termination efficiency. *Annu.Rev.Cell Dev.Biol.*, Vol.15, pp. 661-703.
- Shoemaker, C.J., Eyler, D.E., & Green, R. (2010) Dom34:Hbs1 promotes subunit dissociation and peptidyl-tRNA drop-off to initiate no-go decay. *Science*, Vol.330, pp. 369-372.
- Shyr, Y.Y., Hepburn, A.G., & Widholm, J.M. (1992) Glyphosate selected amplification of the 5-enolpyruvylshikimate-3-phosphate synthase gene in cultured carrot cells. *Mol.Gen.Genet.*, Vol.232, pp. 377-382.
- Sonenberg, N. & Dever, T.E. (2003) Eukaryotic translation initiation factors and regulators. *Curr Opin.Struct.Biol.*, Vol.13, pp. 56-63.
- Spring, J. (1997) Vertebrate evolution by interspecific hybridisation - are we polyploid? *FEBS Lett.*, Vol.400, pp. 2-8.

- Stark, H., Rodnina, M.V., Wieden, H.J., Zemlin, F., Wintermeyer, W., & van Heel, M. (2002) Ribosome interactions of aminoacyl-tRNA and elongation factor Tu in the codon-recognition complex. *Nat.Struct.Biol.*, Vol.9, pp. 849-854.
- Sturtevant, A.H. (1925) The Effects of Unequal Crossing over at the Bar Locus in *Drosophila*. *Genetics*, Vol.10, pp. 117-147.
- Taylor, J.S. & Raes, J. (2004) Duplication and divergence: the evolution of new genes and old ideas. *Annu.Rev.Genet.*, Vol.38, pp. 615-643.
- Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalariao, R.R., Bhalariao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroove, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J.C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de, P.Y., & Rokhsar, D. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, Vol. 313, pp. 1596-1604.
- Urbero, B., Eurwilaichitr, L., Stansfield, I., Tassan, J.P., Le Goff, X., Kress, M., & Tuite, M.F. (1997) Expression of the release factor eRF1 (Sup45p) gene of higher eukaryotes in yeast and mammalian tissues. *Biochimie*, Vol.79, pp. 27-36.
- Valente, L. & Kinzy, T.G. (2003) Yeast as a sensor of factors affecting the accuracy of protein synthesis. *Cell Mol.Life Sci.*, Vol.60, pp. 2115-2130.
- Valle, M., Sengupta, J., Swami, N.K., Grassucci, R.A., Burkhardt, N., Nierhaus, K.H., Agrawal, R.K., & Frank, J. (2002) Cryo-EM reveals an active role for aminoacyl-tRNA in the accommodation process. *EMBO J.*, Vol. 21, pp. 3557-3567.
- van den Elzen, A.M., Henri, J., Lazar, N., Gas, M.E., Durand, D., Lacroute, F., Nicaise, M., van Tilbeurgh, H., Seraphin, B., & Graille, M. (2010) Dissection of Dom34-Hbs1 reveals independent functions in two RNA quality control pathways. *Nat.Struct.Mol.Biol.* Vol.17, pp. 1446-1452.
- van Hoof, A., Frischmeyer, P.A., Dietz, H.C., & Parker, R. (2002) Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science*, Vol.295, pp. 2262-2264.
- van Rijk, A. & Bloemendal, H. (2003) Molecular mechanisms of exon shuffling: illegitimate recombination. *Genetica*, Vol.18, pp. 245-249.
- Vasudevan, S., Peltz, S.W., & Wilusz, C.J. (2002) Non-stop decay-a new mRNA surveillance pathway. *Bioessays*, Vol.24, pp. 785-788.

- Vision, T.J., Brown, D.G., & Tanksley, S.D. (2000) The origins of genomic duplications in *Arabidopsis*. *Science*, Vol.290, pp. 2114-2117.
- Wallrapp, C., Verrier, S.B., Zhouravleva, G., Philippe, H., Philippe, M., Gress, T.M., & Jean-Jean, O. (1998) The product of the mammalian orthologue of the *Saccharomyces cerevisiae* *HBS1* gene is phylogenetically related to eukaryotic release factor 3 (eRF3) but does not carry eRF3-like activity. *FEBS Lett.*, Vol.440, pp. 387-392.
- Wang, W., Czaplinski, K., Rao, Y., & Peltz, S.W. (2001) The role of Upf proteins in modulating the translation read-through of nonsense-containing transcripts. *EMBO J.*, Vol.20, pp. 880-890.
- Wang, Y., Chai, B., Wang, W., & Liang, A. (2010) Functional characterization of polypeptide release factor 1b in the ciliate *Euplotes*. *Biosci.Rep.*, Vol.30, pp. 425-431.
- Wang, Y. & Gu, X. (2000) Evolutionary patterns of gene families generated in the early stage of vertebrates. *J.Mol.Evol.*, Vol.51, pp. 88-96.
- Wolfe, K. (2000) Robustness-it's not where you think it is. *Nat.Genet.*, Vol.25, pp. 3-4.
- Wong, S., Butler, G., & Wolfe, K.H. (2002) Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc.Natl.Acad.Sci.U.S.A.*, Vol.99, pp. 9272-9277.
- Zhouravleva, G., Schepachev, V., Petrova, A., Tarasov, O., & Inge-Vechtomov, S. (2006) Evolution of translation termination factor eRF3: Is GSPT2 generated by retrotransposition of GSPT1's mRNA? *IUBMB.Life.*, Vol.58, pp. 199-202.

# Analysis of Duplicate Gene Families in Microbial Genomes and Application to the Study of Gene Duplication in *M. tuberculosis*

Venu Vuppu and Nicola Mulder

*Computational Biology Group, Department of Clinical Laboratory Sciences  
Institute of Infectious Diseases and Molecular Medicine, Health Science Faculty  
University of Cape Town  
South Africa*

## 1. Introduction

Though considerable sequence information from different organisms was available prior to the recent advances in genome sequencing technology, the foundation for our current understanding of the mechanisms of bacterial pathogenesis was laid by the release of the first complete genome sequence of *Haemophilus influenza* in 1995 (Fraser-Liggett, 2005). Ever since, significant progress in the availability of data for different genomes has been possible due to the contribution of various genome sequencing projects (Koonin & Wolf, 2008). Despite the complete genome sequences of many pathogenic organisms being available, the mortality rates due to these infectious agents still remains a problem, highlighting the need to decipher the complex molecular mechanisms responsible for survival of the bacteria. The wealth of complete genome information for pathogens can be effectively explored using comparative genomic tools for the identification of common and unique sets of genes involved in the propagation of virulence. Sequence comparison tools have been developed to identify homologous genes from the complete genomes of microorganisms. Homologous genes which arise from speciation tend to maintain functions similar to that of their ancestral molecule and are known as orthologs, while the genes originating from duplication events often evolve new functions and are defined as paralogs (Tatusov *et al.*, 1997).

The world of microbes is highly diverse with genome complexity differing across a wide range of microorganisms. In general, the difference in the complexity of genomes is dictated by the life style and environment of the organism (Cordero & Hogeweg, 2009). Life style plays an important role in regulating the genome dynamics of an organism, and functional novelty provided by gene duplication is thought to enhance the adaptation capability of the organism. In addition, horizontal transfer of operons or functional units of genes from external sources may provide an immediate functional benefit to the organism, thereby adding to the functional complexity of the genomes. The availability of complete genome sequences of important mycobacteria such as *Mycobacterium tuberculosis*, *Mycobacterium ulcerans*, *Mycobacterium bovis*, *Mycobacterium leprae*, *Mycobacterium paratuberculosis*, *Mycobacterium avium* and others, can be used to gain deeper insights into possible

mechanisms of prokaryotic gene innovation. Genome data has been used in recent studies to compare different species of mycobacteria, as well as different strains, to understand the evolution and pathogenesis of *M. tuberculosis* (Marri *et al.*, 2006). In our study, duplicate gene sets from different mycobacteria were investigated to identify the distribution of important functional classes of protein families, and the evolution of these functional classes was further analyzed by comparing their genetic divergence following duplication.

The importance of gene duplication in prokaryotic gene innovation is well established and comparative analysis of duplicate genes with basic characteristic features of genomes like GC content and genome size may aid in deciphering their contributions. In contrast to eukaryotes, GC content varies widely across different bacterial genomes (Mann & Chen, 2010), and analysis of GC variations between related bacteria could be useful in establishing evolutionary relationships (Mann *et al.*, 2010). The focus of the majority of earlier studies was on deciphering the role of GC composition in HGT (Nelson *et al.*, 1999; Hamady *et al.*, 2006), transcription start and stop sites (Zhang *et al.*, 2004), nucleotide substitution rates (DeRose-Wilson & Gaut, 2007), optimal growth temperature (Basak & Ghosh, 2005; Musto, 2006) and metabolic characteristics (Naya *et al.*, 2002). Furthermore, genome size has been reported to increase with an increase in number of genes in duplicate gene families (Snel *et al.*, 2002; Pushker *et al.*, 2004). In this study we analyzed the genomes of 56 pathogenic and 20 non-pathogenic microorganisms to identify and characterize the expanded gene families across these organisms. In addition to the GC content, we investigated the relationship between genome size and duplicate gene percentage. On finding sufficient evidence for a correlation between genome size and extent of gene duplication, we further investigated the significance of duplicate genes in enhancing genome complexity. He and Zhang (2005) previously reported the importance of gene duplication in enhancing genome and organism complexity in eukaryotes. However, due to the difference in the selective pressures operating on prokaryotic and eukaryotic genomes, we used the duplicate and single copy genes to investigate the influence of protein lengths on genome and organism complexity of prokaryotic organisms, with a specific focus on investigating the role of duplicate genes in enhancing the genome complexity of *M. tuberculosis*.

## 2. Materials and methods

### 2.1 Data selection and identification of homologous sequences

Comparative sequence analysis of different genomes is the most common approach for identifying orthologs and paralogs. However, here we used both the sequence and protein signature data as the latter could substantiate the former, and enables identification of more distantly related members of a protein family. We collected non-redundant protein sets for 76 microorganisms, including pathogens and non-pathogens, to identify expanded gene families in these organisms. The selection of the non-pathogenic bacteria in this study is of value, since many of these may also contain virulent genes which could act as barriers conferring protection against the defense mechanisms of the host, thus enhancing the survival capabilities and adapting the organism to intracellular conditions. In addition, acquisition of specific virulent gene clusters can transform these non-pathogenic agents to pathogenic microorganisms.

For the selected organisms, approximately 1,91,497 protein signatures, 2,47,858 protein sequences, Genome size and G+C composition data were retrieved from the InterPro (<http://www.ebi.ac.uk/interpro>) (Apweiler *et al.*, 2001; Mulder *et al.*, 2007) and Integr8 (<http://www.ebi.ac.uk/integr8>) (Kersey *et al.*, 2005) databases respectively. The protein

signature data from InterPro enabled the identification of approximately 27,827 proteins which exhibited complete domain identity (same InterPro matches) over their entire length to one or more proteins in *M. tuberculosis* strain H37Rv. Within each organism and across all organisms, the proteins showing complete domain identity were grouped together as duplicate gene sets or ortholog and paralog sets, respectively, and those with no common signature matches were considered to be single copies. In addition to the identification of expanded families using InterPro data, homologous sequences were clustered using BlastClust in two separate clustering procedures:

- a. **Independent Genome Clustering:** This involves within genome clustering to generate clusters of paralogs or protein families for each genome. BlastClust was executed at a wide range of percentage identities over varying lengths of the sequence to select the optimum parameters. Amongst the tested parameters, a 30% similarity over 60% sequence length cut-off was chosen, as it generated a suitable number of clusters (in line with previously reported numbers of duplicated families for *M. tuberculosis*).
- b. **Multiple Genome Clustering:** In this, all of the 76 genomes were appended together for the clustering of related proteins (orthologs and paralogs). In addition, the clustering of six of the mycobacterial species was performed separately for the evolutionary analysis of expanded gene families in *M. tuberculosis*.

## 2.2 Evolutionary analysis

For evolutionary studies, in addition to 66 paralogous gene clusters, 116 multiple genome clusters from the phylogenetic matrix of six of the closely related mycobacterial genomes that showed gene family expansions in both *M. tuberculosis* and *M. leprae*, as well as other mycobacteria, were selected. The proteins in each of the clusters were aligned with T-coffee (Notredame *et al.*, 2000), and poorly aligned regions were edited using the Gblocks program (Castresana, 2000) with adjustments in the default settings for the generation of optimal sequence alignments. For each of these protein alignments, selection of the best-fitting amino acid substitution model was performed according to the Akaike informational criterion, and the gamma correction factor ( $\alpha$ ), the proportion of invariable sites (I), and observed amino acid frequencies (F) were estimated and selected for subsequent phylogenetic analysis using ProtTest (Abascal *et al.*, 2005). Since, PhyML is a maximum likelihood method with the ability to incorporate the estimated values of  $\alpha$ , proportion of invariable sites, and observed frequencies, the tree topologies for the gene sets in the identified clusters were constructed using this program (Guindon & Gascuel, 2003). The genetic distance measures from each of the estimated tree topologies were used to compute average and maximum genetic distance using Perl scripts.

## 3. Results and discussion

### 3.1 Identification of expanded gene families and relation to GC content and genome size

We used sequence clustering and protein signature data to identify expanded genes families within and across several different microbial genomes. The across-genome clustering of protein sequence data yielded 1,984 expanded genes in 441 clusters for *M. tuberculosis* H37Rv. The protein signature method allowed us to group 30,885 proteins into 2238 clusters from all the organisms. InterPro signatures usually match between 50% and 80% of a genome, so data is not available for every protein. Since signature data enables identification

of more distantly related members of a cluster, but loses data where proteins do not match InterPro, the sequence and signature-based cluster data was merged, and used to generate a phylogenetic profile, which reflected the number of copies in each expanded family for each organism. From this, we identified 2011 duplicate/expanded genes in 461 clusters for *M. tuberculosis*, confirming previous reports that the duplicate genes make up approximately half of the *M. tuberculosis* genome (Tekaia *et al.*, 1999). The percentages derived from the 2 methods are shown in Table 1 for the 6 mycobacteria studied. The 461 clusters in the merged data include gene families that are also expanded in different organisms.

S.No	Organism	Sequence	Signature	Union
1	<i>M. tuberculosis</i>	31.47%	38%	50.96%
2	<i>M. bovis</i>	30.28%	42%	48.69%
3	<i>M. paratuberculosis</i>	39.75%	49%	56.46%
4	<i>M. avium</i>	42.06%	49%	55.19%
5	<i>M. ulcerans</i>	36.82%	46%	53.51%
6	<i>M. leprae</i>	12.03%	20%	30.44%

Table 1. Percentage of the genome belonging to expanded gene families for the mycobacteria. Data was generated using sequence clustering, protein signatures and a combination of the two (union).

Next, we investigated the GC composition of different bacteria in relation to the duplicate gene percentages (Figure 1) to understand the characteristic features of genomes maintaining high percentages of duplicate genes. A statistical analysis of the data using Pearson's correlation, revealed a moderate correlation between the GC content and estimated duplicate gene percentages. However, the analysis of the trend lines of the scatter plot in figure 1 reveals the presence of three different kinds of relationships in the data: i) an initial increase of the trend line, ii) the initial increase is followed by a phase of neutrality, iii) and a steady increase of the trend line follows the phase of neutrality. We analyzed histograms of GC content and duplicate gene percentages and observed differences in the modality of the data distribution; GC percentages followed a trimodal distribution, while the duplicate gene percentages followed a unimodal distribution. Thus, although a positive correlation could be inferred from the analysis of the scatter plot, the correlation coefficient could be subdued by the differences in the modality of the data distributions. Hence, based on the analysis of the scatter plot and trimodal distribution of the GC percentage histogram, the organisms were grouped into three categories based on their GC compositions:

**Group 1:** Organisms having GC content greater than 54 percent.

**Group 2:** Organisms having GC content greater than 44 percent and less than 54 percent.

**Group 3:** Organisms having GC content less than 44 percent.

We then performed a one-way ANOVA on the data (Table 2). Taking into consideration the mean square values (Mean Sq) and the calculated p-value of  $2 \times 10^{-16}$ , we predicted that the mean variance between groups is significant compared to the within sample variance. These results indicate the existence of differences in the means of the three groups of organisms, and hence, we reject the null hypothesis and accept the alternative. Further, to estimate how significantly different the means of each group are compared to one another, a Tukey's Honest Significant Difference (Tukey's HSD) test was performed, and significant differences in the mean values of group2 and group1, group3 and group1, and group3 and group2 were found. From the results table of the Tukey's multiple comparison test (Table 3), it can be



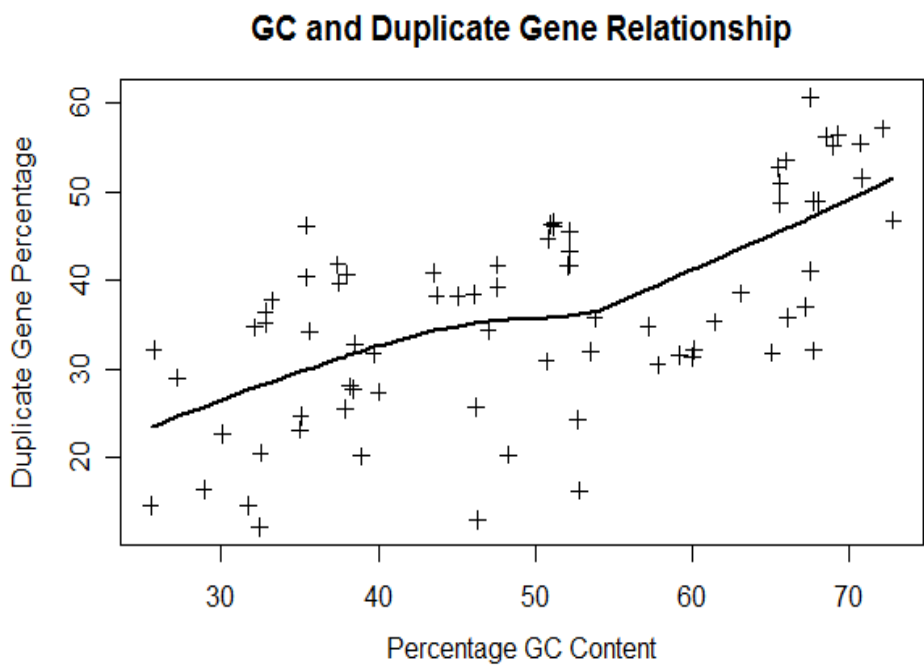


Fig. 1. Scatter plot analysis of the relationship between GC content and duplicate gene percentages of the selected organisms. The percent GC content is plotted on the X-axis and duplicate gene percentage on the Y-axis. The graph suggests that a positive correlation between GC content and duplicate gene percentages exists for the majority of the investigated organisms.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Groups	2	13174.8	6587.4	417.28	<2.2e-16***
Residuals	73				
Signif Codes	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 2. One-Way ANOVA Results. The columns of the table display the degrees of freedom (df), sum square values (Sum Sq), Mean square values (Mean Sq), F value and p-value (Pr(>F)) reported by One-Way ANOVA for the data.

Tukey's Multiple Comparison of Means			
Groups	Diff	Lwr	Up
G2-G1	-16.36	-19.08	-13.64
G3-G1	-31.54	-34.15	-28.92
G3-G2	-15.18	-17.88	-12.48

Table 3. The table displays the differences between the mean values of the groups. The Groups column represents the investigated groups: group2 and group1 (G2-G1), group3-group1 (G3-G1), and group3-group2 (G3-G2). The differences in the means of the groups are given by the difference (diff) column, and the lower (lwr) and upper (upr) columns represent the lower and upper boundaries for the estimated mean difference between the groups.

inferred that both groups, G2-G1 and G3-G2 exhibit similar mean differences (-16.36 and -15.18). However, the mean differences of both these groups are higher than the mean difference (-31.54) of group G3-G1. Therefore, group2 organisms, which have higher mean differences compared with group1 and group3, could be responsible for the reduced correlation coefficient values. Hence, their elimination from the list of investigated organisms could result in the prediction of strong positive correlation between the GC composition and duplicate gene percentages of group1 and group3 organisms. Thus, we suggest that gene duplication events may be a characteristic feature of GC rich bacterial genomes. Since all of the selected mycobacterial species in the present study are representatives of group1, the phenomenon of gene duplication in this genus could be attributed to its high GC content.

In addition to GC compositions, we analyzed the influence of duplicate genes on the physical expansion of the genomes (Figure 2). An observed correlation coefficient value of 0.84 at a p-value of  $2 \times 10^{-16}$  between the genome size and duplicate genes provides sufficient evidence to prove the contribution of duplicate genes to genome expansion of these organisms. This is not surprising, since the addition of genes through gene duplication will obviously increase genome size unless some genes are lost in the process.

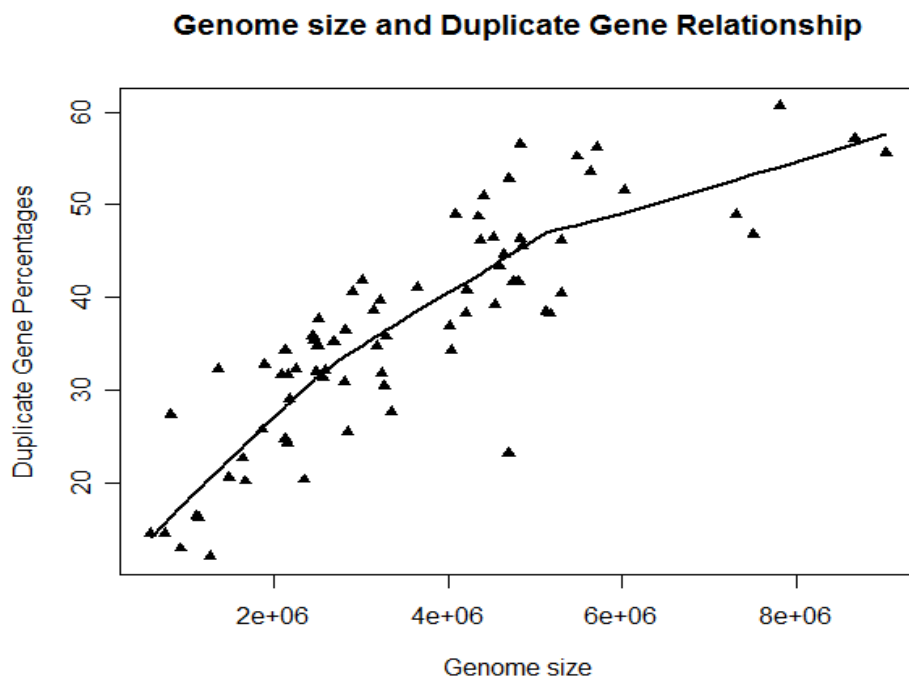


Fig. 2. The graph displays the relationship between duplicate gene percentage and genome size for the selected organisms. The identified duplicate gene percentages are plotted on the X-axis and genome size on the Y-axis. From the graph, a positive correlation can be observed between duplicate gene percentages and genome size.

We went on to investigate the domain complexity of single and duplicate copy genes to further enhance our understanding of the functional complexity of these organisms. As a measure of domain complexity, the number of domains present in each of the

corresponding proteins of the genes was computed from InterPro data, and the mean for the total number of domains was estimated for the duplicate and single copy genes using Perl scripts. From figure 4, we can see that the mean number of domains per duplicate gene is lower than that of the single copy genes. This suggests that single copy genes should be more complex due to the presence of more domains. We further analyzed the results, by statistically comparing the difference in the mean domain numbers of duplicate and single copy genes using the Mann-Whitney U test. From the resulting W value of 5717 at a p-value of  $2 \times 10^{-16}$ , we inferred that the mean number of domains per single copy genes is significantly higher than that of duplicate genes. Thus, these studies suggest that the single copy genes are functionally more complex than duplicate genes. This was a surprising result, given that the mean length of duplicate genes was found to be higher than that of the single copy genes. Therefore, we specifically investigated the influence of gene lengths on the domain complexity of *M. tuberculosis*, and compared this statistic in two other organisms, *Leptospira interrogans* and the model organism, *Escherichia coli*.

Functional complexity of Duplicate and Single Copy Genes

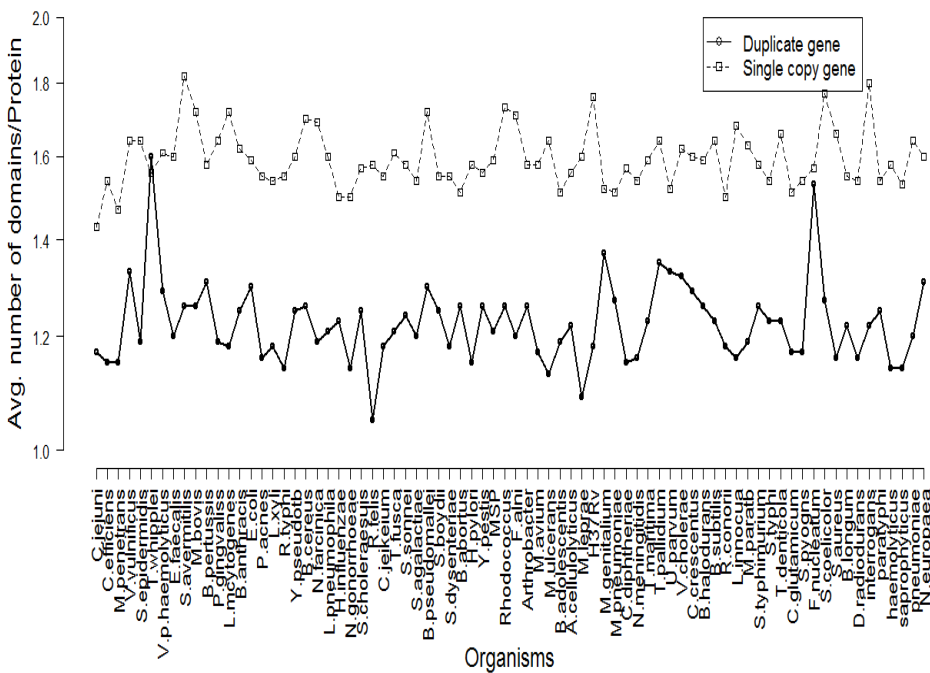


Fig. 4. Comparison of Functional Complexity of the Expanded and Single Copy Genes Based on InterPro Signature Data in Selected organisms. The graph displays the mean number of domains per protein in the duplicate and single copy genes of each organism. The organisms investigated are plotted on the X-axis and the corresponding mean number of domains per protein for each organism on the Y-axis.

For each of these three organisms, the total number of genes in the genome was retrieved, and for every gene, we estimated the total number of domains to determine the relationship between gene length and domain number (Figure 5 -Whole Genome Analysis). In addition, the number of domains for each of the duplicate (Figure 6) and single copy genes (Figure 7) were also estimated. The preliminary analysis of the relationships using scatter plots suggested that the number of domains per gene does not necessarily increase with an increase in the gene length. Further, correlation coefficient values estimated from the Pearson moment correlation were used for statistical confirmation of the relationships.

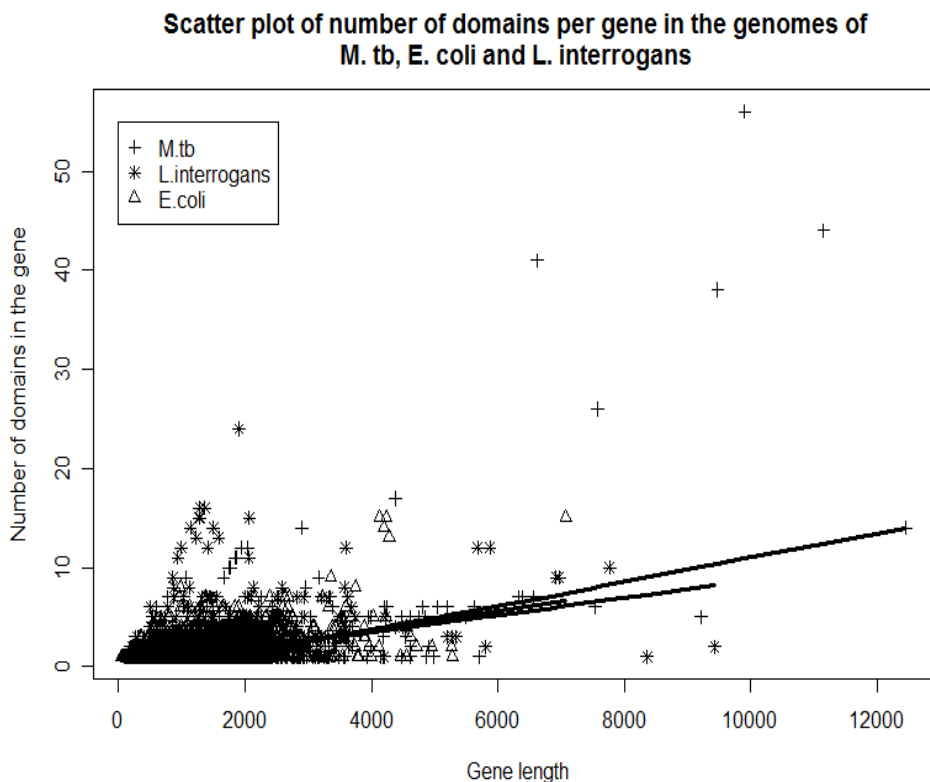


Fig. 5. Investigation of number of domains per gene in the *M. tuberculosis* H37Rv, *E. coli* and *L. interrogans* genomes (Whole Genome Analysis). The graph displays the relationships between gene length and number of domains. The sequence lengths of each gene are plotted on the X-axis and the corresponding number of domains per gene on the Y-axis. From the graph, it can be inferred that the gene length is not necessarily dependent on domain number.

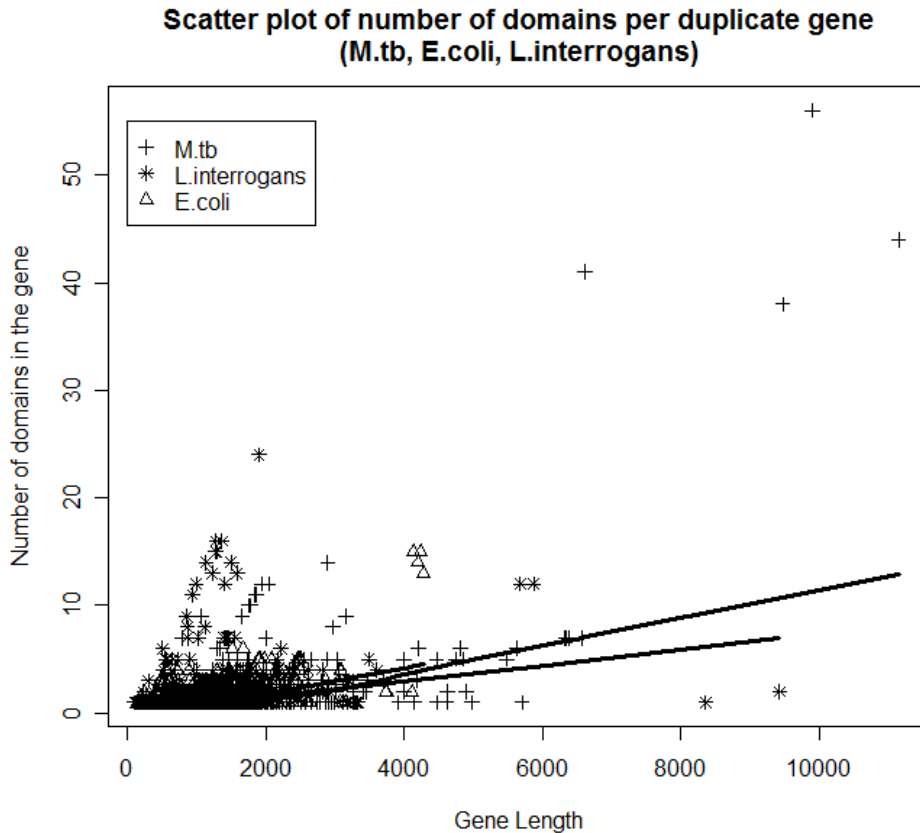


Fig. 6. Investigation of number of domains per duplicate gene in the *M. tuberculosis* H37Rv, *E. coli* and *L. interrogans* genomes (Duplicate Gene Analysis). The graph displays the relationship between sequence length (X-axis) and number of domains (Y-axis) of the duplicate genes.

The correlation coefficient values for the whole genome, duplicate gene and single copy gene analysis in *E. coli* were 0.48, 0.47, and 0.49, respectively, while the values of 0.39, 0.25, and 0.53 were reported for *L. interrogans* (Table 4). The results from these two organisms suggest that the number of domains does not increase significantly with the increase in gene length and hence, domain complexity may be independent of the gene length or vice versa. Although the reported correlation coefficient values of 0.58, 0.62 and 0.59 corresponding to the whole genome, duplicate and single copy gene analyses, respectively, in *M. tuberculosis* are higher than those for *E. coli* and *L. interrogans*, these correlation coefficient values still do not suggest significant positive correlation between domain complexity and gene lengths. Thus, the specific protein complexity studies of these three genomes show that it is not necessarily surprising that while the duplicate genes are generally longer than single copy genes, they tend to contain fewer domains.

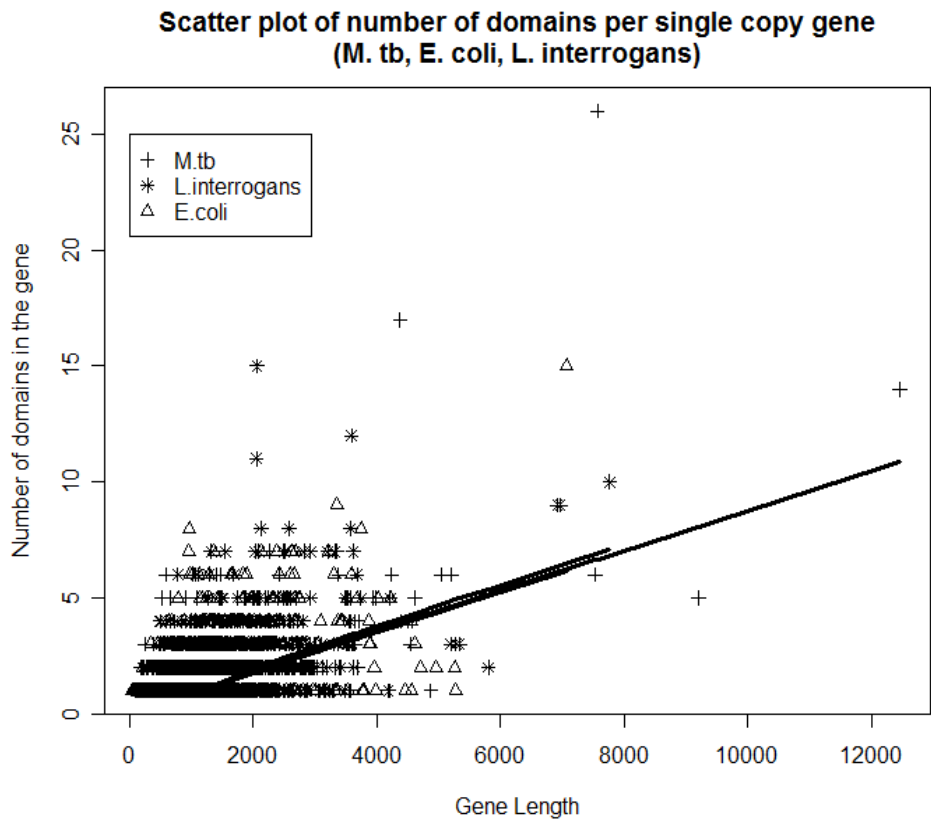


Fig. 7. Investigation of number of domains per single copy gene in the *M. tuberculosis* H37Rv, *E. coli* and *L. interrogans* genomes (Single Copy Gene Analysis). The graph displays the relationship between sequence length and number of domains of the single copy genes in these genomes.

Pearson's product-moment correlation						
Organism and P-value	<i>E. coli</i>	P-value	H37Rv	P-Value	<i>L. interrogans</i>	P-Value
All Proteins	0.48	2.2e-16	0.58	2.2e-16	0.39	2.2e-16
Duplicate	0.47	2.2e-16	0.62	2.2e-16	0.25	2.64e-10
Single	0.49	2.2e-16	0.59	2.2e-16	0.53	2.2e-16

Table 4. Correlation Coefficient and P-values of the whole genome, duplicate and single copy gene analysis in *E. coli*, *M. tuberculosis* H37Rv and *L. interrogans*. The table displays the results of Pearson's product-moment correlation.

### 3.3 Functional and evolutionary analysis of expanded genes in *M. tuberculosis*

The protein sequence and signature data for the 76 genomes were clustered into related sets of duplicate genes for the study of relationships between percentage duplication and the GC content, genome size and gene complexity described above. However, since the comparison of closely related organisms is better for inferring evolutionary relationships, we separately clustered six of the closely related mycobacterial genomes and identified 390 duplicate gene clusters in *M. tuberculosis*. The results were represented as a phylogenetic profile and a summary is shown in Table 5.

Organism	Total Genes	TGC	SCG	DGC	Total duplicate Genes	Estimated Duplicate Gene Percentages
<i>M. tuberculosis</i>	3947	2815	2425	390	1521	38.53
<i>M. bovis</i>	3910	2817	2439	378	1471	37.62
<i>M. paratuberculosis</i>	4316	2807	2343	464	1973	45.71
<i>M. avium</i>	5040	3199	2679	520	2361	46.84
<i>M. ulcerans</i>	4206	2755	2359	396	1847	43.91
<i>M. leprae</i>	1036	1603	1261	119	342	21.33

Table 5. Protein sequence clustering of the mycobacterial group. The columns of the table represent the selected organisms, total number of genes in the genome, total number of identified gene clusters (TGC), total number of single copy genes (SCG), total number of duplicate gene clusters (DGC) in the organism, total number of duplicate genes in the duplicate gene clusters identified, and the percentage of duplicate genes estimated for each organism.

The biggest expanded family in *M. tuberculosis* was the PE/PPE/PGRS family with 164 members, followed by a family of alcohol dehydrogenases and oxidoreductases with 44 members, the fatty-acid-CoA ligase family with 33 members, then acyl-CoA dehydrogenase with 27 members. A manual assignment of high-level functional classes was done previously in the laboratory for all *M. tuberculosis* proteins. This was used here to determine the functional distribution of all 390 expanded families in this organism. Figure 8 shows the number of families and number of proteins belonging to each of the functional classes. The biggest class is made up of enzymes or proteins involved in metabolism, followed by proteins of unknown function. From the data, we selected 116 gene clusters which showed gene family expansions in *M. tuberculosis* and *M. leprae*, as well as other mycobacteria. We are interested in expansion in *M. leprae*, as this is a highly reduced mycobacterial genome, so expanded genes that have been maintained are likely to be important. When considering only the 116 families that are also expanded in *M. leprae*, the distribution of functional classes is similar, except for a large reduction in the number of unknown protein families.

For each of the 116 clusters of interest, we calculated the genetic distance between family members and investigated the relationship between genetic distance and gene family size. Our results suggest that the genetic distance between two of the most distant proteins in the clusters increases with an increase in cluster size. In addition, the correlation coefficient value of 0.87 at a p-value of  $2.2 \times 10^{-16}$  is indicative of strong positive correlation between these factors. These sets of duplicate copies of proteins are clustered from different mycobacterial genomes, and since the estimated maximum genetic distance between the 2



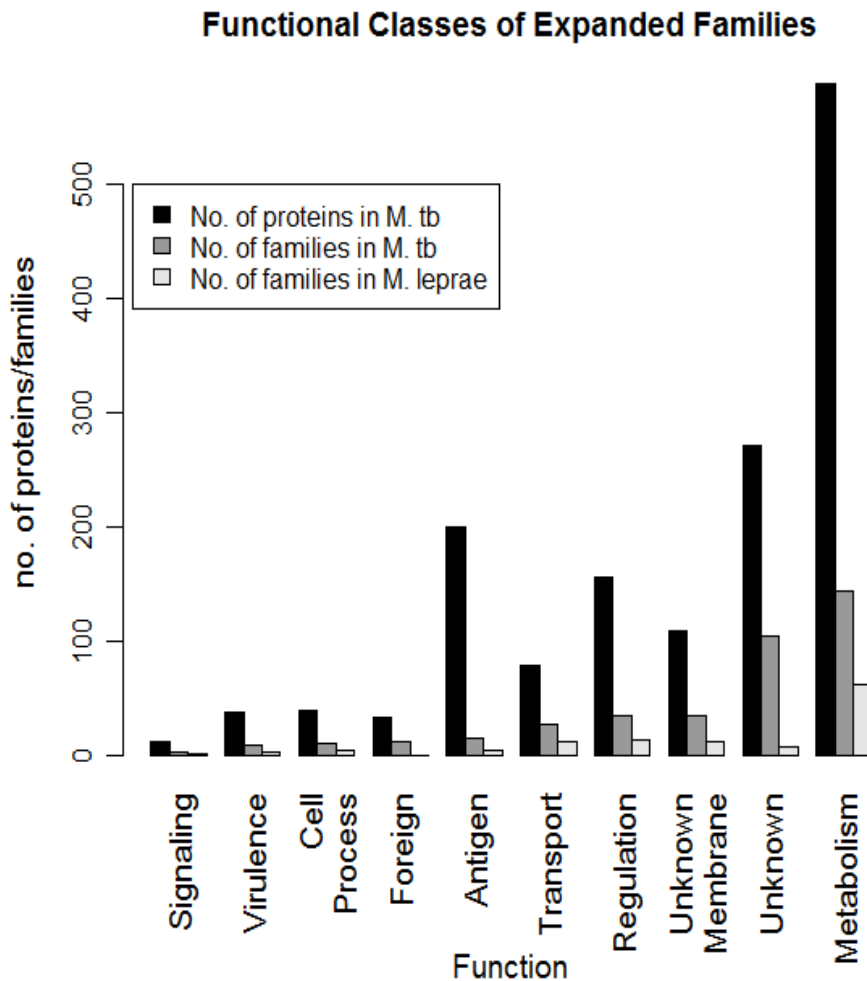


Fig. 8. Distribution of functions in *M. tuberculosis* (all 390 clusters) and *M. leprae*-shared (116 clusters) expanded families.

most distant proteins in each of these sets increases with an increase in cluster size, it was inferred that some of the duplicate copies show a tendency to diverge from the original ancestral functions after multiple duplication events in bigger families. To investigate the average divergence of proteins in these clusters, the relationship between average genetic distance and cluster size was determined. The results suggest that the average genetic distance between the gene families does not increase with the cluster size, except perhaps for the few larger families. In order to statistically verify the results, correlation coefficient values were estimated using the Pearson's product-moment correlation. The correlation coefficient value of 0.43 at a p-value of  $1.17 \times 10^{-6}$  indicates the presence of moderate

correlation between the average genetic distance and cluster size. It suggests that the majority of homologous gene families identified from these mycobacterial species have not undergone significant functional divergence and still show close evolutionary relatedness, but these results may be skewed by the fact that the clusters contain orthologs and paralogs. Orthologs are generally predicted to maintain similar functions, while paralogs are known to have diverged functions.

The relationship between cluster size and genetic distance was also studied for 66 paralogous families only (within genome clustering). Within each of the selected mycobacterial species, the estimated tree topologies were used to investigate the genetic divergence of the identified paralogous gene families by computing the maximum genetic distance between two of the most distant paralogs in each of the clusters. A scatter plot analysis of the computed maximum genetic distances and cluster sizes was performed (Figure 9), and correlation coefficient values were estimated for studying the genetic divergence of these paralog gene families. From the analysis of the scatter plot (Figure 9) and correlation coefficient values (Table 6), we inferred that the genetic distance between the two most distant proteins increases with the cluster size.

Organism	df	Pearson's correlation	P-value
<i>M. tuberculosis</i>	64	0.88	2.20e-16
<i>M. bovis</i>	62	0.81	4.4e-16
<i>M. paratuberculosis</i>	59	0.93	2.20e-16
<i>M. avium</i>	56	0.95	2.20e-16
<i>M. ulcerans</i>	59	0.93	2.20e-16
<i>M. leprae</i>	36	0.66	5.94e-06

Table 6. Results of the correlation calculations for maximum genetic distance versus cluster size, including degrees of freedom (df), Pearson's correlation coefficient values, and the corresponding p-values.

In addition to maximum genetic distance, the average genetic distance for each of the paralog gene families was computed to investigate the evolutionary relationships between the members within the selected mycobacterial genomes. To provide statistical significance for the scatter plot observations, correlation coefficient values were estimated using the Pearson's product-moment correlation (Table 7). The scatter plot (Figure 10) and correlation coefficient values (Table 7), suggest a moderate negative correlation between the average genetic distance and cluster size of the paralogous gene families.

### 3.4 Further analysis of one example expanded gene family in *M. tuberculosis*

While we have evolutionary data for all the orthologous and paralogous families of *M. tuberculosis*, we cannot show all the results, so we have selected an important class of regulatory proteins as an example. The adaptability of *M. tuberculosis* to enable successful survival of the stressful conditions in the host during infection is attributed to the existence of a diverse class of sigma factors in the organism (Fontan, 2009). The organism is suggested to contain numerous sigma factors that bind to the core subunit of RNA polymerase to provide promoter specificity (Fontan, 2008). To investigate the phylogenetic diversification of sigma factors in *M. tuberculosis* and other mycobacteria, we studied the sigma factors

Organism	df	Pearson's correlation	P-value
<i>M. tuberculosis</i>	64	-0.44	0.0001966
<i>M. bovis</i>	62	-0.5	2.38e-05
<i>M. paratuberculosis</i>	59	-0.41	0.0007649
<i>M. avium</i>	56	-0.43	0.0007819
<i>M. ulcerans</i>	59	-0.43	0.000633
<i>M. leprae</i>	36	-0.44	0.005978

Table 7. Pearson’s correlation coefficient results for the relationship between the average genetic distance and cluster size. The columns include degrees of freedom (df), Pearson’s correlation coefficient values, and the corresponding P-values.

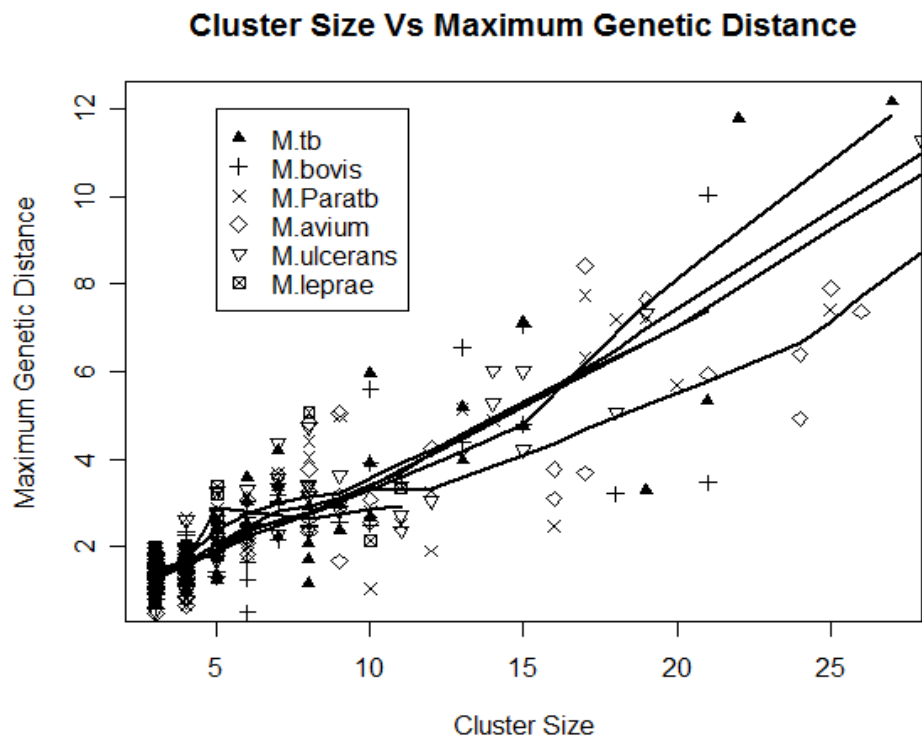


Fig. 9. Relationship between maximum genetic distance and cluster size for families of *M. tuberculosis* H37Rv, *M. bovis*, *M. paratuberculosis*, *M. avium*, *M. ulcerans* and *M. leprae*. The X-axis represents the cluster size (total proteins in each cluster) and Y-axis shows the genetic distance between the two most distant proteins in the clusters of each organism. The genetic distance appears to increase with the cluster size, suggesting a correlation between them.

identified by our ortholog and paralog clustering methods. From the analysis of the sigma factor phylogenetic trees of *M. tuberculosis* (Figure 11), we infer that gene duplication events followed by divergence could have resulted in the bifurcation of the sigma factor class of proteins into two subfamilies (marked as A and B in the Figure).

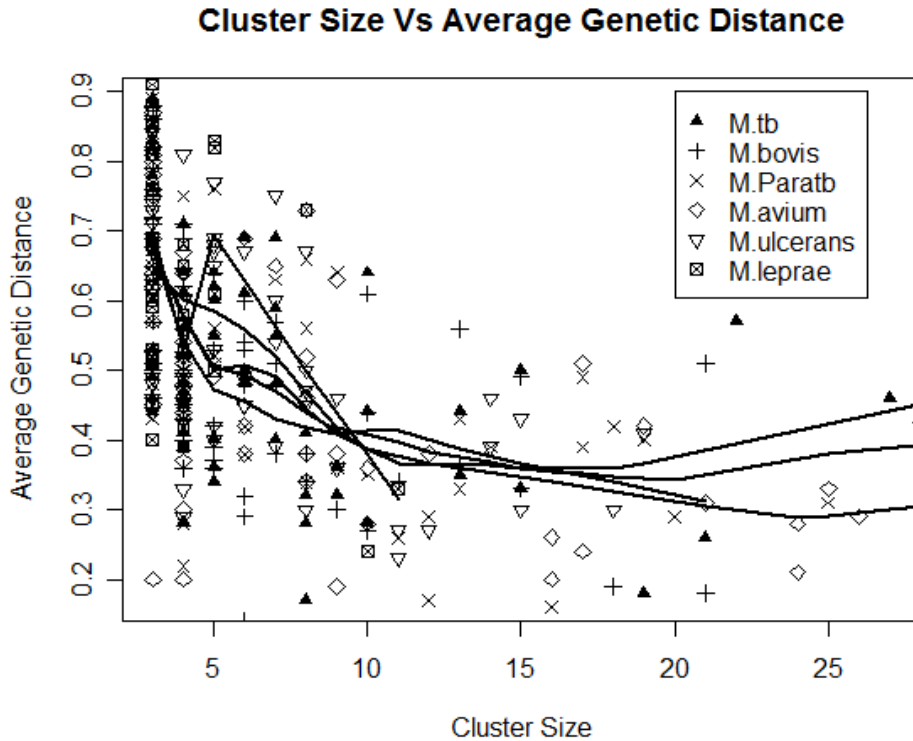


Fig. 10. Relationship between average genetic distance and cluster size for duplicate gene families of *M. tuberculosis* H37Rv, *M. bovis*, *M. paratuberculosis*, *M. avium*, *M. ulcerans* and *M. leprae*. The X-axis represents the cluster size (total proteins in each cluster) and Y-axis shows the average genetic distance between the identified gene families of each organism. The average genetic distance appears to decrease with the cluster size, suggesting a negative correlation between these two factors.

### 3.4.1 Analysis of sigma factor proteins in subfamily A

Following duplication, the proteins of this subfamily have diverged into 2 groups: SigE and SigM (Figure 11). The 2 proteins in the SigE group have further diverged following duplication and divergence. However, one of the proteins in the sigE group was identified to have no orthologs in *M. leprae* (Figure 12), and loss of various sigma factors is suggested to be the reason for *M. leprae* reductive genome evolution (Babu, 2003). Interestingly, all the paralogs of *M. tuberculosis* appear to have orthologs in *M. bovis*, but the absence of sigM proteins in *M. bovis*, and the large divergence of this protein group in *M. tuberculosis* compared to other mycobacteria enables us to speculate on its significance in *M. tuberculosis* evolution. Though an error in available *M. bovis* sequences could have resulted in incorrect annotation of the sigM locus as a pseudogene (Manganelli *et al.*, 2004), the extent of divergence of this protein in *M. tuberculosis* compared to other mycobacteria prompts further investigation into its possible paths of pseudogenization or neofunctionalization.

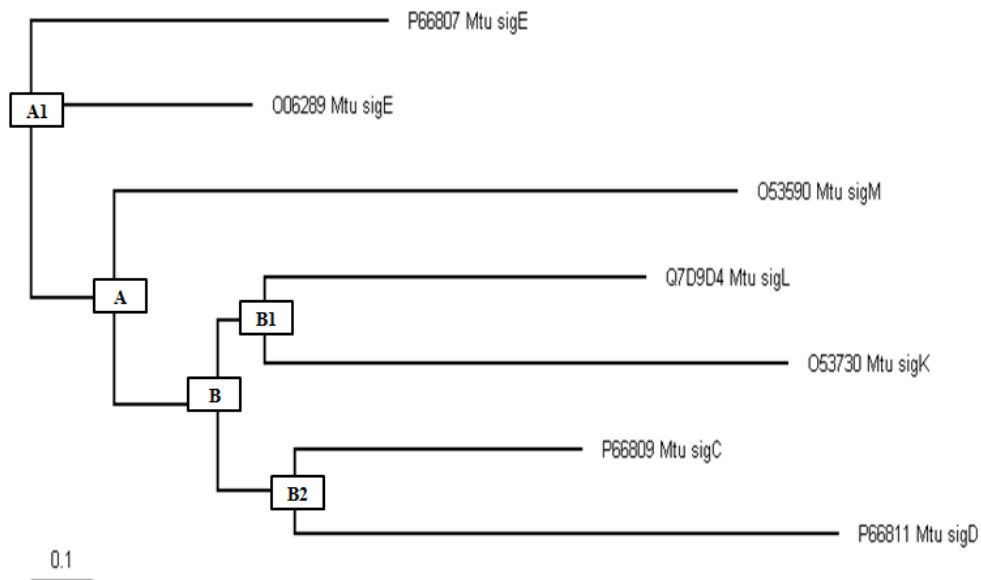


Fig. 11. Phylogenetic tree of the Sigma factor paralogs cluster inferred by the maximum likelihood method. The duplication events are marked by A's and B's. The figure displays the phylogenetic diversification of sigma factors (sigE, sigM, sigL, sigK, sigC and sigD).

### 3.4.2 Analysis of the sigma factor proteins in subfamily B

From the analysis of the sigma factor class of paralogs in GroupB (Figure 11), we infer that duplication events followed by divergence resulted in the 2 groups of sigma factor subfamilies (marked as B1 and B2 in Figure 11). The sigma factor proteins in each of the subfamilies have significantly diverged after gene duplication. For the two sigma proteins (sigK and sigL) in one of the subfamilies (Figure 12), sigK was identified to have no orthologs in *M. avium*, *M. paratuberculosis* or *M. leprae*, and sigL was noted to have no orthologs in *M. leprae*. These results are inconsistent with the published reports of Manganelli *et al*, 2003. For the other subfamily of sigma proteins (sigC and sigD) on the phylogenetic tree (Figure 12), we did not identify orthologs for sigC in *M. paratuberculosis* or sigD in *M. leprae*.

## 4. Discussion

The availability of complete genome sequences of many bacteria and significant progress in the development of modern computational biology methods has resulted in the evolution of a powerful platform for the comparative investigation of genome diversity across different organisms. Here, we make use of the wealth of genome information and bioinformatics tools to understand the significance of gene duplication in *M. tuberculosis* evolution. The investigation of relationships between the GC composition and duplicate gene percentages identified from the sequence and InterPro domain data provides sufficient evidence to suggest a positive correlation between them for group1 and group3 organisms. Here, the mycobacterial species are part of the group1 organisms, so the maintenance of

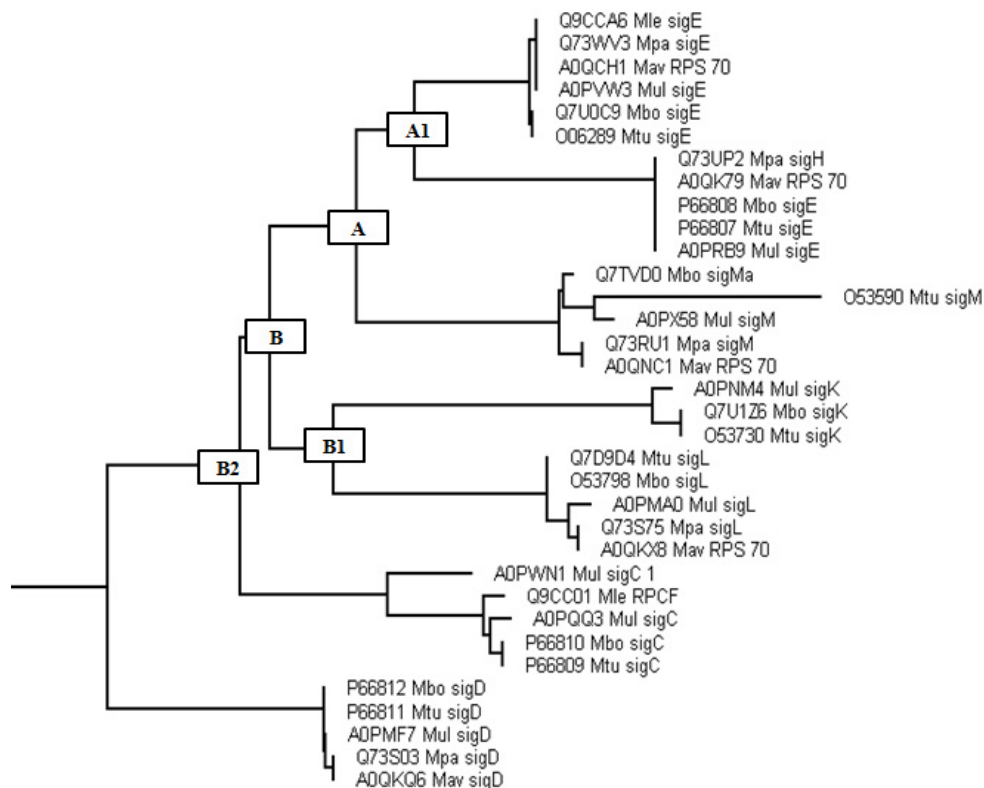


Fig. 12. Phylogenetic tree of the Sigma factor ortholog and paralog cluster inferred by the maximum likelihood method. The duplication events are marked by A's and B's. The labels Mtu, Mbo, Mav, Mul, Mpa and Mle represent proteins from *M. tuberculosis*, *M. bovis*, *M. avium*, *M. ulcerans*, *M. paratuberculosis* and *M. leprae* respectively.

high duplicate gene percentages in these species, with the exception of *M. leprae*, could be attributed to the high GC composition of their genomes. Unsurprisingly, the study has also shown a correlation between duplicate gene percentage and genome size, suggesting that gene duplication increases genome size. Further, our investigations on protein complexity provide deeper insights into the general trend in gene length and domain number in duplicate genes in these organisms.

He and Zhang (2005), investigating *Saccharomyces cerevisiae*, showed duplicate genes to be complex molecules with longer sequences containing more functional domains. From the investigations of the mean gene lengths of 76 pathogenic and non-pathogenic organisms, it is evident that the average length of the duplicate genes is comparatively higher than that of single copy genes. However, the analysis of mean number of domains in the duplicate and single copy genes reveals the presence of a higher number of domains in the single copy genes compared to the duplicate genes. According to Stoltzfus (1999), partial loss-of-function

mutations lead to the preservation of duplicate genes with single functions (Stoltzfus, 1999; Lynch & Force, 2000). Moreover, it has been suggested that duplicate genes lose one of the domains that were originally present in the ancestral molecule, and by complementation of the lost domains, both the daughter copies are reported to reflect the original ancestral function. Thus, gene complexity is suggested to be reduced after subfunctionalization of duplicate genes (He & Zhang, 2005). Further, the complementary loss of subfunctions is considered to facilitate the preservation of duplicate gene pairs, and due to relaxed evolutionary constraints following subfunctionalization, the chances of long-term evolution of new functions is enhanced (Force *et al.*, 1999). However, since deleterious mutations are more common than beneficial mutations (Cun, 2010), evolution of new and essential protein functions is considered to be a rare event (Nadeau & Sankoff, 1997; Force *et al.*, 1999). According to the predominant argument, the evolution of new domains would be favored only if they can perform a function different to that of preexisting domains or domain combinations (Lagomarsino *et al.*, 2009). Further, the majority of duplicate genes are predicted to develop new functions from the already existing ancestral gene functions, and if new functions evolve by mutation from prior domains, it is less likely that all of the domains would evolve into new domains due to the mutational bridge for new domain evolution being too far from the ancestral molecule (Lagomarsino *et al.*, 2009). Hence, evolution of new functions following subfunctionalization could be a rare event. Therefore, the presence of fewer domains in the duplicate genes compared to the single copy genes could be due to evolution of duplicate genes by subfunctionalization, where complementary loss of subfunctions is viewed to primarily facilitate preservation of the duplicate gene. Alternatively, the addition of new domains into a bacterial genome could be due to acquisition by HGT. Indeed, acquisition of one or more domains by HGT in 30 to 50 percent of bacteria has been reported (Choi & Kim, 2007). The acquired gene or gene segment is known to be beneficial only if it has some properties different to that of recipient genome (Kinsella *et al.*, 2003). Since the selection of a new domain would depend upon its ability to perform a biological function that is not covered by pre-existing domains, addition of such rare domains by HGT could be an uncommon phenomenon (Lagomarsino *et al.*, 2009). Adaptation of bacteria to new environments requires evolution of new functions (Hooper & Berg, 2003), and gene duplication is viewed to be the general mechanism of adaptation to different environmental conditions (Kondrashov, 2002). However, a recent study suggests HGT to be a far more important route to adaptation compared to gene duplication (Koonin & Wolf, 2009). Further, duplication of horizontally transferred genes with weak or no functions is suggested to accelerate the evolutionary process of gene innovation. Since both gene duplication and HGT are considered to be important routes of bacterial adaptation to changing environments, and amplification of weak ancillary functions is considered to be the easiest route to gene innovation, quick adaptation of bacteria to changing environments could be due to amplification of weak ancillary functions. Thus, reduced functional complexity of the investigated duplicate genes compared to single copy genes could be due to preservation of the majority of the paralogs by subfunctionalization, and the rare event of neofunctionalization could have been either due to divergence of subfunctions over an evolutionary period of time following preservation of subfunctionalized paralogs, or mostly due to rapid amplification of weak ancillary functions after gene duplication. To gain deeper insights into the functional complexity of duplicate genes in *M. tuberculosis*, we focussed on the evolutionary analysis of the duplicate genes in six of the closely related mycobacterial species.

From the analysis of maximum genetic distance between the two most distant proteins of the mycobacterial multiple genome clusters, we suggest that the divergence of at least one of the duplicate gene copies from the ancestral gene increases with the increase in cluster size. These homologous gene families consist of orthologs and paralogs. The lack of a strong correlation between the average genetic distance and cluster size of the duplicate gene copies in the multiple genome clusters indicates that the homologous gene families including proteins from different mycobacterial species have not undergone complete functional divergence. This is to be expected for orthologs, which tend to maintain their functions.

The average genetic distance estimated for single genome paralogous gene clusters, on the other hand, decreases with the increase in cluster size, suggesting that, on average, smaller families tend to diverge more rapidly than the larger families. This is apart from some members of the larger families, which have obviously diverged further as they are contributing to the increased maximum genetic distance with cluster size. Though gene duplication is considered to be an important mechanism for acquiring new genes, and creating evolutionary novelty (Torgerson and Singh, 2004), horizontal gene transfer (HGT) is also known to be a wide spread phenomenon, and a significant proportion of genes in bacteria are accepted to have been acquired by HGT (Price *et al.*, 2007). The genome of *M. tuberculosis* is known to contain 19 genes of eukaryotic origin, and it is speculated that the organism may have also acquired genes from other prokaryotes by HGT (Kinsella *et al.*, 2003). In addition, the occurrence of many intraspecies HGT events in the progenitor of *M. tuberculosis* has been reported (Rosas-Magallanes *et al.*, 2006). The ability of HGT to incorporate a new gene which is homologous to an existing gene family member is well recognized (Ochman, 2001; Kinsella *et al.*, 2003; Krzywinska, 2004), and in comparison to its gene family members, the newly introduced gene may be more divergent in sequence and function (Pushker *et al.*, 2004). Following duplication, such laterally transferred genes with already divergent functions may further diversify in the process of evolving new functions, and this could result in an increase in genetic distance between the laterally transferred duplicate gene and its paralog gene family members. The chance of this should increase with the number of members.

Phylogenetic analysis of the sigma factors in *M. tuberculosis* suggested that most of the sigma factors have orthologs in other mycobacteria. However, we could not observe orthologs in *M. leprae* for a few of the subfamilies, and this could have been due to the extensive loss of sigma factors during its reductive genome evolution. Agarwal *et al.*, 2007 reports that sigM proteins control only a small subset of genes, and their loss would not influence *M. tuberculosis* virulence (Agarwal *et al.*, 2007). The difference in the divergence of sigM in *M. tuberculosis* compared to other mycobacteria, and absence of its ortholog in *M. bovis* should be considered further to study the importance of the sigM factor in *M. tuberculosis* virulence.

## 5. Conclusions and future work

The estimated duplicate gene percentages for *M. tuberculosis* from independent genome clustering (31%), InterPro signature methods (38%), across genome clustering (49%) and a union of the methods (51%) were all relatively high, showing the significance of gene duplication in *M. tuberculosis* genome evolution. The investigation of relationships between the GC composition and duplicate gene percentages identified from the sequence and InterPro domain data provides sufficient evidence to suggest that for the mycobacterial species, with the exception of *M. leprae*, the maintenance of high duplicate gene percentages



could be attributed to the high GC composition of their genomes. The study has also shown a correlation between duplicate gene percentage and genome size, suggesting that gene duplication increases genome size, which is a logical result. Interestingly, our functional complexity results were in contrast to recent finding in eukaryotes, and we show that, on average, duplicate genes have longer sequences but fewer domains than single copy genes in the investigated organisms. The reduced functional complexities of duplicate genes could be due to their evolution by subfunctionalization following duplication.

We also show that duplicate gene families of mycobacterial multiple genome clusters have not undergone complete functional divergence following gene duplication and still tend to maintain their functions. Our maximum genetic distance results suggest that multiple duplication events in a few of the duplicate copies of bigger families may result in their functional divergence from the original ancestral functions. Our paralog maximum genetic distance results suggest that the increase in genetic distance between the two most distant proteins with the size of the gene family may be due to duplication followed by paralog evolution of some of the distant genes that already have divergent functions compared to its paralog members. From the study of average genetic distance of paralogs, we suggest that slow evolution of paralogs of large families in *M. tuberculosis* could be due to preservation of original ancestral functions by the mechanism of subfunctionalization.

For future studies, we are investigating selection pressure and comparison between smaller and larger gene family evolution to shed light on the evolutionary fate of duplicate genes and functional innovation in *M. tuberculosis*. In addition, since the functional constraints on amino acid residues are known to differ due to the potential changes in protein function, we have studied site specific rate differences between the amino acids of closely related mycobacterial species to aid in deciphering specific subfamily evolutionary divergence following gene duplication. Such predicted critical amino acids when mapped on to protein secondary structure could help in evaluation of important structural locations in functional diversification. In addition to functional divergence, gene expression data from different experimental conditions is of use to understand the degree of expression divergence of the genes following duplication events. Overall, we are working on further investigation of *M. tuberculosis* duplicate genes with the integration of phylogeny-sequence-structure-function-expression information, which will be valuable for understanding the functional and evolutionary fate of genes following gene duplication in *M. tuberculosis*.

## 6. Acknowledgement

We thank the National Bioinformatics Network and Computational Biology Group, University of Cape Town, South Africa for supporting this work.

## 7. References

- Abascal, F.; Zardoya, R. & Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics Applications Note*, Vol. 21, No. 9, pp. 2104–2105.
- Agarwal, N.; Woolwine SC, Tyagi S, Bishai WR. (2007). Characterization of the *Mycobacterium tuberculosis* Sigma Factor SigM by Assessment of Virulence and Identification of SigM-Dependent Genes. *Infection and Immunity*, Vol. 75, No 1, pp. 452–461.

- Apweiler, R.; Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, Vol. 29, No. 1, pp. 37-40.
- Babu, MM. (2003). Did the loss of sigma factors initiate pseudogene accumulation in *M. leprae*? *Trends in Microbiology*, Volume 11, No. 2, pp. 59-61.
- Basak, S. & Ghosh, TC. (2005). On the origin of genomic adaptation at high temperature for prokaryotic organisms. *Biochemical and Biophysical Research Communications*, Vol. 330, No. 3, pp. 629-632.
- Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, Vol. 17, Vol. 4, pp. 540-552.
- Choi, I. & Kim, S. (2007). Global extent of horizontal gene transfer. *Proceedings of National Academy of Science*, Vol. 104, No. 11, pp. 4489-4494.
- Cordero, OX. & Hogeweg, P. (2009). The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proceedings of National Academy of Science*, Vol. 106, No. 51, pp. 21748-21753.
- Cun Y. (2010). The Evolutionary Dynamics of Mutant Allele at Duplicate Loci *arXiv:1007.0333v1*.
- DeRose-Wilson, LJ. & Gaut, BS. (2007). Transcription-related mutations and GC content drive variation in nucleotide substitution rates across the genomes of *Arabidopsis thaliana* and *Arabidopsis lyrata*. *BMC Evolutionary Biology*, Vol. 7, No. 66.
- Fontan, PA.; Voskuil MI, Gomez M, Tan D, Pardini M, Manganello R, Fattorini L, Schoolnik GK, Smith I. (2009). The Mycobacterium tuberculosis Sigma Factor B Is Required for Full Response to Cell Envelope Stress and Hypoxia In Vitro, but It Is Dispensable for In Vivo Growth. *Journal of Bacteriology*, Volume 191, No. 18, pp. 5628-5633.
- Fontan, PA.; Aris V, Alvarez ME, Ghanny S, Cheng J, Soteropoulos P, Trevani A, Pine R, and Smith I. (2008 ). Mycobacterium tuberculosis Sigma Factor E Regulon Modulates the Host Inflammatory Response. Vol. 198, pp. 877-85.
- Force, A.; Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, Vol. 151, No. 4, pp. 1531-1545.
- Fraser-Liggett, CM. (2005). Insights on biology and evolution from microbial genome sequencing. *Genome Research*, Vol. 15, No. 12, pp.1603-1610.
- Guindon, S. & Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, Vol. 52, No. 5, pp. 696-704.
- Hamady, M; Betterton, MD., & Knight, R. (2006). Using the nucleotide substitution rate matrix to detect horizontal gene transfer. *BMC Bioinformatics*, Vol. 7, No. 476.
- He, X. & Zhang, J. (2005). Gene Complexity and Gene Duplicability. *Current Biology*. Vol. 15, No. 11, pp. 1016-1021.
- Hooper, DS. & Berg, GO. (2003). On the Nature of Gene Innovation: Duplication patterns in Microbial Genomes. *Molecular Biology and Evolution*, Volume 20, No. 6, pp. 945-954.

- Kersey, P.; Bower L, Morris L, Horne A, Petryszak R, Kanz C, Kanapin A, Das U, Michoud K, Phan I, Gattiker A, Kulikova T, Faruque N, Duggan K, McLaren P, Reimholz B, Duret L, Penel S, Reuter I and Apweiler R. (2005). Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Research*, Vol. 33, Database issue, D297-D302.
- Kinsella, RJ.; Fitzpatrick DA, Creevey CJ, McInerney JO. (2003). Fatty acid biosynthesis in *Mycobacterium tuberculosis*: Lateral gene transfer, adaptive evolution, and gene duplication. *Proceedings of National Academy of Science*, Vol. 100, No. 18, pp. 10320-10325.
- Kondrashov, FA.; Rogozin IB, Wolf YI, Koonin EV. (2002). Selection in the evolution of gene duplications. *Genome Biology*, Vol. 3, No. 2.
- Koonin, EV. & Wolf YI. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, Vol. 36, No. 21.
- Koonin, EV. & Wolf, YI. (2009). Is evolution Darwinian or/and Lamarckian? *Biology Direct*, Vol. 4, No. 42.
- Krzywinska, E.; Krzywinski J, & Schorey JS. 2004. Naturally occurring horizontal gene transfer and homologous recombination in *Mycobacterium*. *Microbiology*, Vol. 150, No. Pt 6, pp. 1707-1712.
- Lagomarsino, MC. (2009). Universal features in the genome-level evolution of protein domains. *Genome Biology*, Vol. 10, No. 1.
- Manganelli, R.; Provvedi R, Rodrigue S, Beaucher J, Gaudreau L, Smith I. (2004).  $\sigma$  Factors and Global Gene Regulation in *Mycobacterium tuberculosis*. *Journal of Bacteriology*, Volume 186, No. 4, pp. 895-902.
- Marri, PR.; Bannantine, JP. & Golding, GB. (2006). Comparative genomics of metabolic pathways in mycobacterium species: gene duplication, gene decay and lateral gene transfer *FEMS. Microbiology Review*, Vol. 30, No. 6, pp. 906-925.
- Mulder, NJ; Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. (2007). New developments in the InterPro database. *Nucleic Acids Research*. Vol. 35, No. D224-D228.
- Musto, H.; Naya H, Zavala A, Romero H, Alvarez-Valin F, and Bernardi G. (2006). Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochemical and Biophysical Research Communications*, Vol. 347, No. 1, pp. 1-3.
- Mann, S. & Chen, YP. (2010). Bacterial genomic G + C composition-eliciting environmental adaptation. *Genomics*, Vol. 95, No. 1, pp. 7-15.
- Mann, S.; Li, J. & Chen YP. (2010). Insights into Bacterial Genome Composition through Variable Target GC Content Profiling. *Journal of Computational Biology*, Vol. 17, No. 1, Pages 79-96.
- Nadeau, JH. & Sankoff, D. (1997). Comparable Rates of Gene LOSS and Functional Divergence After Genome Duplications Early in Vertebrate Evolution. *Genetics*, Vol. 147, No. 3, pp. 1259-1266.

- Naya, H.; Romero H, Zavala A, Alvarez B, Musto H. (2002). Aerobiosis Increases the Genomic Guanine Plus Cytosine Content (GC%) in Prokaryotes. *Journal of Molecular Evolution*, Vol. 55, No. 3, pp. 260-264.
- Nelson, KE.; Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM. (1999). Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritime*. *Nature*, Vol. 399, pp. 323-329.
- Notredame, C.; Higgins, DG. & Heringa J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, Vol. 302, No. 1, pp. 205-217.
- Ochman, H. 2001. Lateral and oblique gene transfer. *Current Opinion in Genetics & Development*, Vol. 11, No. 6, pp. 616-619.
- Price, MN.; Dehal PS, & Arkin AP. 2007. Orthologous transcription factors in bacteria have different functions and regulate different genes. *PLoS Computational Biology*, Vol. 3, No. 9, pp. 1739-1750.
- Pushker, R.; Mira A, & Rodriguez-Valera F. (2004). Comparative genomics of gene-family size in closely related bacteria. *Genome Biology*, Vol. 5, No. 4.
- Rosas-Magallanes, V.; Deschavanne P, Quintana-Murci L, Brosch R, Gicquel B, Neyrolles O. 2006. Horizontal Transfer of a Virulence Operon to the Ancestor of *Mycobacterium tuberculosis*. *Molecular Biology and Evolution*, Vol. 23, No. 6, pp. 1129-1135.
- Snel. B.; Bork, P. & Huynen MA. (2001). Genomes in Flux: The Evolution of Archaeal and Proteobacterial. *Gene Content Genome Research*, Vol 12, No. 1, pp.17-25.
- Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. *Journal of Molecular Biology*, Vol 49, No. 2, pp. 169-181.
- Tatusov, RL.; Koonin, EV. & Lipman DJ. (1997). A Genomic Perspective on Protein Families. *Science*, Vol. 278, No. 631.
- Tekaia, F.; Gordon SV, Garnier T, Brosch R, Barrell BG, Cole ST. (1999). Analysis of proteome of *mycobacterium tuberculosis* in silico. *Tubercle and Lung Diseases*, Vol. 79, No. 6, pp. 329-342.
- Torgerson, DG. & Singh RS. 2004. Rapid Evolution Through Gene Duplication and Subfunctionalization of the Testes-Specific  $\alpha 4$  Proteasome Subunits in *Drosophila*. *Genetics*, Vol. 168, No. 3, pp. 1421-1432.
- Zhang, L.; Kasif S, Cantor CR, Broude NE. (2004). GC/AT-content spikes as genomic punctuation marks. *Proceedings of National Academy of Science*, Vol. 101, No. 48, pp. 16855-16860.

# The Evolutionary History of CBF Transcription Factors: Gene Duplication of CCAAT – Binding Factors NF-Y in Plants

Alexandro Cagliari, Andreia Carina Turchetto-Zolet, Felipe dos Santos Maraschin, Guilherme Loss, Rogério Margis and Marcia Margis-Pinheiro  
*Universidade Federal do Rio Grande do Sul/UFRGS  
 Brazil*

## 1. Introduction

Eukaryotic gene expression is often controlled by complex and refined combinatorial transcription factor networks composed of multiprotein complexes that derive their gene regulatory capacity from both intrinsic properties and from their *trans*-acting partners (Singh, 1998; Wolberger, 1998; Remenyi *et al.*, 2004). Participation in such higher complex order allows an organism to use single transcription factors to control multiple genes with different temporal and spatial expression patterns (Siefers *et al.*, 2009).

In this chapter, we provide a synopsis of the genetic and genomic mechanisms that might be responsible for the gene copy diversification observed in the eukaryotic NF-Y transcription factor family. We identify the genes coding for NF-Y transcription factors in eukaryotes with an emphasis on the duplication of the NF-Y family in the plant lineage and discuss the important consequences of its gene diversification.

## 2. The CCAAT *cis*-element promoter

Eukaryotic genes contain numerous *cis*-regulatory elements that mediate their induction, repression or basal transcription (Dyan and Tjian, 1985; Myers *et al.*, 1986; Maity and de Crombrughe, 1998). These regulatory elements can be found in the proximity of transcribed genes, such as the promoter region and/or in distant regions of the genes where they may act as enhancers (de Silvio *et al.*, 1999).

The transcriptional regulation of several eukaryotic genes is coordinated through sequence-specific binding of proteins to the promoter region located upstream of the gene. During evolution, many of these protein-binding sequences, which are found in a wide variety of organisms, have shown a high degree of conservation (Edwards *et al.*, 1998).

The CCAAT box is one of the most common upstream elements, found in approximately 25–30% of eukaryotic promoters (Bucher, 1990; Mantovani, 1998). It is typically located between 60–100 bp upstream of the transcription start site and it can function in direct or in inverted orientations (Dorn *et al.*, 1987b; Bucher, 1990; Edwards *et al.*, 1998; Mantovani, 1998; Stephenson *et al.*, 2007) with possible cooperative interactions between multiple boxes (Tasanen *et al.*, 1992) or other conserved motifs (Muro *et al.*, 1992; Rieping and Schoffl, 1992;

Edwards *et al.*, 1998). CCAAT boxes are highly conserved within homologous genes across species in terms of position, orientation, and flanking nucleotides (Mantovani, 1998). In addition, the spacing between the CCAAT box and other promoter-specific *cis*-elements is also conserved among species (Dorn *et al.*, 1987a; Chodosh *et al.*, 1988; Maity and de Crombrughe, 1998). The expression of genes under the control of promoters that contain CCAAT boxes may be ubiquitous or tissue/stage specific, suggesting that the gene expression pattern is also determined by other *cis* and *trans* elements (Stephenson *et al.*, 2007).

In *Sacharomyces cerevisiae*, CCAAT boxes are found in the promoters of cytochrome genes, in genes coding for proteins that are activated by non-fermentable carbon sources (McNabb *et al.*, 1995) and in genes involved in nitrogen metabolism (Dang *et al.*, 1996). In the filamentous fungus *Aspergillus nidulans*, CCAAT boxes are present in genes involved with penicillin biosynthesis (Steidl *et al.*, 1999). In higher eukaryotes, a multitude of promoters contain CCAAT boxes, including those of developmentally controlled and tissue-specific genes (Berry *et al.*, 1992), housekeeping and inducible genes (Roy and Lee, 1995) and cell-cycle regulated genes (Mantovani, 1998). In addition, many cell-cycle regulated promoters lack a recognizable TATA-box, but contain more than one CCAAT box in a position close to and sometimes overlapping with the start site of transcription (Zwicker and Muller, 1997).

### 3. The CBF/NF-Y transcription factor

Several CCAAT-binding proteins have been isolated and described, including CBF/NF-Y (CCAAT Binding Factor/Nuclear Factor of the Y box), CTF/NF1 (CCAAT Transcription Factor/Nuclear Factor 1), C/EBP (CCAAT/Enhancer Binding Protein) and CDP (CCAAT Displacement Protein) (Mantovani, 1999). Among them, NF-Y is the most ubiquitous and specific one acting as a key proximal promoter factor in the transcriptional regulation of an array of different eukaryotic genes. Unlike other CCAAT-binding proteins, NF-Y requires a high degree of conservation of the CCAAT pentanucleotide sequence and shows strong preference for specific flanking sequences (Dorn *et al.*, 1987a; Stephenson *et al.*, 2007). Therefore, the NF-YC transcription factor can be distinguished from the other CCAAT-binding proteins based on its DNA sequence requirements (Maity and de Crombrughe, 1998).

The CBF/NF-Y transcription factor, which will be referenced in this chapter as NF-Y, is a conserved oligomeric transcription factor found in all eukaryotes that is involved in the regulation of diverse genes (Maity *et al.*, 1992; McNabb *et al.*, 1995; Edwards *et al.*, 1998; Mantovani, 1998; Siefers *et al.*, 2009). NF-Y typically acts in concert with other regulatory factors to modulate gene expression in a highly controlled manner (Nelson *et al.*, 2007). In many eukaryotic promoters, the functional NF-Y-binding sites are relatively close to the TATA motif (Bucher, 1990) and are invariably flanked by at least one additional functionally important *cis*-element. Several reports have shown that various factors, including transcription factors, co-activators, and TATA-binding proteins, interact with NF-Y or its subunits in promoting transcriptional regulation (Mantovani, 1999; Yazawa and Kamada, 2007). NF-Y was originally identified as the protein that recognizes the MHC class II conserved Y box in Ea promoters (Dorn *et al.*, 1987a; Matuoka and Chen, 2002). It specifically recognizes the consensus sequence 5'-CTGATTGGYYRR-3' or 5'-YYRRCCAATCAG-3' (Y is 5 pyrimidines and R is 5 purines) present in the promoter region of eukaryotic genes. Bioinformatic analyses indicate that about 30% of mammalian promoters have predicted

NF-Y binding sites (Bucher, 1990; Testa *et al.*, 2005), and chromatin immunoprecipitation data have demonstrated additional widespread NF-Y binding in nonpromoter sites.

Suggesting the importance of binding context, NF-Y-regulated gene expression can be tissue specific, developmentally regulated, or constitutive (Maity and de Crombrughe, 1998; Siefers *et al.*, 2009). The transcriptional activity of NF-Y can be regulated by differential expression, alternative splicing, protein-protein interactions, and cellular redox potential (Matuoka and Yu Chen, 1999).

NF-Y has been shown to be involved in the regulation of some G1/S genes whose expressions are attenuated during the senescence process (Matuoka and Yu Chen, 1999). NF-Y plays a pivotal role in the cell cycle regulation of the mammalian cyclin A, *cdc25C*, and *cdc2* genes, in the S-phase of the cell cycle (Currie, 1998). Additionally, there are a number of genes involved in the cellular response to damage and stress, including the phospholipid hydroperoxide glutathione peroxidase genes (Huang *et al.*, 1999), which are regulated by NF-Y, indicating its pivotal role in the removal of damaging agents from cells (Matuoka and Chen, 2002). Although NF-Y functions basically as a transactivator of gene expression, it is also involved, directly or indirectly, in the downregulation of transcription. For instance, NF-Y binds to the mouse CCAAT box renin enhancer and blocks the binding of positive regulatory elements (Shi *et al.*, 2001). In this case, NF-Y dysfunction would lead to the damage of systems that control blood pressure (Matuoka and Chen, 2002).

NF-Y is composed of three different subunits named NF-YA (also known as HAP-2 or CBF-B), NF-YB (HAP3 or CBF-A), and NF-YC (HAP5 or CBF-C) that interact to form a complex that can bind CCAAT DNA motifs and control the expression of target genes (Figure 1). Each subunit is required for DNA binding, subunit association and transcriptional regulation in both vertebrates and plants (Sinha *et al.*, 1995; Stephenson *et al.*, 2007). Yeast possesses a fourth subunit, called HAP4, which provides a transcriptional activation domain to the complex (Forsburg and Guarente, 1989; Lee *et al.*, 2003). The yeast HAP4 protein is not needed for DNA-binding but contains an acidic domain that is essential to promote transactivation when associated with the HAP2/HAP3/HAP5 complex (Olesen and Guarente, 1990; Serra *et al.*, 1998). In vertebrates, the function of this fourth domain was incorporated into other subunits (Forsburg and Guarente, 1989; Yazawa and Kamada, 2007). Despite the wide cellular distribution and functional variability of NF-Y-regulated genes, most eukaryotic genomes have only one or two genes encoding each NF-Y subunit (Maity and de Crombrughe, 1998; Riechmann and Ratcliffe, 2000). Fungi and animals, for example, present single genes encoding each protein subunit. Thus, there is minimal combinatorial diversity in the subunit composition of the heterotrimeric NF-Y in these organisms (Siefers *et al.*, 2009). In contrast, the NF-Y complex in vascular plants is generally encoded by gene families (Riechmann and Ratcliffe, 2000).

### 3.1 NF-Y subunits

NF-Y is the only transcription factor thus far identified for which the interaction of three heterologous subunits creates the DNA binding domain (Maity and de Crombrughe, 1992; McNabb *et al.*, 1995; Sinha *et al.*, 1995). All three NF-Y subunits are essential for the DNA binding activity and one molecule of each subunit forms the NF-Y-DNA complex (Maity and De Crombrughe, 1996). Each NF-Y subunit contains a conserved domain with identities greater than 70% across species. This highly conserved domain is located at the C-terminus of NF-YA; in the central part of NF-YB; and at the N-terminus of NF-YC (Li *et al.*, 1992).

The NF-YA conserved domain can be divided in two functionally distinct regions: an N-terminal region that is required for NF-YB and NF-YC association and a C-terminal region required for DNA-binding (Maity and de Crombrughe, 1992). Additionally, NF-YA usually contains a glutamine (Q)-rich and a serine/threonine (S/T)-rich regions. There are numerous variants of NF-YA due to alternative splicing at the Q-S/T domains (Li *et al.*, 1992) and, although the expression of these isoforms is variable depending of tissue and cell types, they all seem intact in terms of transcriptional function (Matuoka and Chen, 2002).

Both NF-YB and NF-YC subunits possess the highly conserved histone-fold motif (HFM) and are structurally similar to core histone subunits H2B and H2A, respectively, and to the archaeobacterial histone-like protein Hmf-2 (Arents and Moudrianakis, 1995; Baxeavanis *et al.*, 1995; Mantovani, 1998). In terms of identity, NF-YB is 30% identical to H2B, 14% to H2A, 17% to H4 and 18% to H3; NF-YC is 21% identical to H2A, 15% to H4 and H3 and 20% to H2B (Liberati *et al.*, 1999). Other proteins showing a remarkable identity (25-30%) to both NF-YB and NF-YC are present in *Archaea*. These proteins homodimerize and associate with DNA, forming nucleosome-like structures (Sandman *et al.*, 1990). The NF-YB and NF-YC subunits also contain residues that are important for their contact with DNA (Romier *et al.*, 2003; Stephenson *et al.*, 2007). In contrast, the conserved segment of NF-YA has no homology with the histone-fold motif, or with any of the known dimerization motifs present in other heteromeric DNA-binding proteins (Maity and de Crombrughe, 1992).

Some portions of NF-YA, NF-YB and NF-YC present a high degree of identity with yeast HAP3, HAP2 and HAP5, respectively. These HAP genes, which are components of the yeast CCAAT-binding protein, are necessary for the expression of genes encoding components of the electron transport chain. Yeast strains mutated for either of the three genes failed to grow on media containing a nonfermentable carbon source such as lactate or glycerol, a characteristic respiratory-defect phenotype (McNabb *et al.*, 1995).

Assembly of the NF-Y heterotrimer in mammals (where this complex is better studied) follows a strict, stepwise pattern (Sinha *et al.*, 1995; Sinha *et al.*, 1996) (Figure 1). Initially, the NF-YB and NF-YC subunits form a tight heterodimer (Figure 1a) similar to those of the HFM, a conserved protein-protein and DNA-binding interaction module (Luger *et al.*, 1997) composed by 65 amino acid stretch common to all histones that is required for nucleosome formation (Baxeavanis *et al.*, 1995; Luger *et al.*, 1997; de Silvio *et al.*, 1999). This dimer then moves to the nucleus, where the third subunit (NF-YA, Figure 1b) is recruited to generate the complete, heterotrimeric NF-Y (Figure 1c). Interestingly, NF-YA is unable to interact with the NF-YB or NF-YC alone, interacting only with the NF-YB-NF-YC heterodimer (Serra *et al.*, 1998). The complete NF-Y is able to bind promoters containing the core pentamer nucleotide sequence CCAAT (Figure 1d) with high specificity and affinity resulting in either positive or negative transcriptional regulation (Figure 1e) (Peng and Jahroudi, 2002; 2003; Ceribelli *et al.*, 2008; Siefers *et al.*, 2009).

Because the NF-Y transcription factor contains H2B-like and H2A-like molecules (NF-YB and NF-YC, respectively), the complex presents all the core histone components and could mimic the interaction of the nucleosome core with genomic DNA (Struhl and Moqtaderi, 1998). In this scenario, it has been demonstrated that the NF-YA/NF-YB/NF-YC trimer or the NF-YB/NF-YC dimer can bind to H3/H4 tetramer during nucleosome assembly (Caretti *et al.*, 1999). In addition, the NF-Y complex also can bind to the chromatin even after nucleosome formation, indicating the ability of NF-Y to interact with genomic DNA assembled in the nucleosome. The interaction between the NF-Y transcription factor and the DNA molecule causes local disruption of the nucleosomal architecture (Coustry *et al.*, 2001).



This disruption results in a partial dissociation of DNA from the histone core, which might enable the access of the general transcription machinery to initiate the transcription process.

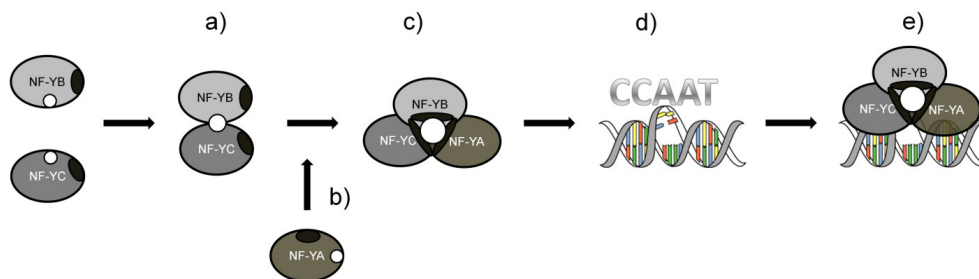


Fig. 1. Assembly of NF-Y subunits and its binding to DNA. Initially, the NF-YB and NF-YC subunits form a tight heterodimer via protein-protein interactions (a). The dimer then moves to the nucleus, where is recruited the third subunit (NF-YA) (b) to generate the complete, heterotrimeric NF-Y (c) that is able to bind promoters containing the core pentamer nucleotide sequence CCAAT (d) resulting in either positive or negative transcriptional regulation (e). Adapted from Mantovani (1999). White circles and oblong black circles into each NF-Y subunit represent the DNA-binding domain and the NF-Y interaction domain of NF-Y subunits, respectively.

#### 4. Gene duplication and evolution

DNA duplication act as one of the main forces driving the evolution of organisms by creating the raw genetic material that natural selection can subsequently modify. Gene duplications arise in eukaryotes at a rate of 0.01 paralogs per gene per million years (Lynch and Conery, 2000), the same order of magnitude of the mutation rate per nucleotide per year (De Grassi *et al.*, 2008). Duplication of individual genes, chromosomal segments, or entire genomes represent the primary source for the origin of evolutionary novelties, including new gene functions and expression patterns (Holland *et al.*, 1994; Sidow, 1996; Lynch and Conery, 2000). However, how duplicated genes successfully evolve from an initial state of complete redundancy, wherein one copy is likely to be expendable, to a stable situation in which both copies are maintained by natural selection, is unclear (Sidow, 1996; Lynch and Conery, 2000; Ober, 2010).

In the evolutionary history of plants, genome duplications have been relatively common, leading to the hypothesis that most angiosperms are to some extent polyploidal (Soltis, 2005). The genome of *Arabidopsis*, for example, possesses traces of at least three polyploidy events (Vision *et al.*, 2000; Simillion *et al.*, 2002), followed by subsequent gene loss (Bowers *et al.*, 2003; Ober, 2010).

Similar to a point mutation, a duplication that occurs in an individual can be fixed or lost in the population. Compared with pre-existing alleles, if a new allele of the duplicate gene is selectively neutral, it has a small probability ( $1/2N$ ) to be fixed in a diploid population (where  $N$  is the effective population size). This suggests that the majority of duplicated genes will be lost. For those duplicated genes that do become fixed, the fixation time averages is  $4N$  generations (Kimura, 1989; Zhang, 2003).

On an evolutionary scale, gene duplication may result in new functions via different scenarios. Although the most likely outcome is a loss of function in one of the two gene copies (nonfunctionalization, Figure 2a), in rare instances one copy may acquire a novel evolutionarily advantageous function and become preserved by natural selection (neofunctionalization, Figure 2b), while the other copy retains the original function. Alternatively, after duplication, mutations may occur in both genes leading to specialization to perform complementary functions (subfunctionalization, Figure 2c) (Lynch and Conery, 2000; Lynch and Force, 2000). This process produces novel genetic variants that drive genetic innovation (Lynch and Conery, 2000; Conrad and Antonarakis, 2007). Because gene duplication generates functional redundancy, it is often not advantageous to the organism to possess two identical genes. In nonfunctionalization (Figure 2a), the accumulation of deleterious mutations might lead to the loss of the original function of one paralogue. Alternatively, instead of being completely lost, many duplicated genes are silenced or become pseudogenes and are thus either unexpressed or functionless (Gallagher *et al.*, 2004; Nicole *et al.*, 2006; Yang *et al.*, 2006; Beisswanger and Stephan, 2008; Xiong *et al.*, 2009). Pseudogenization is the most frequent fate of duplicated genes. In *Caenorhabditis elegans*, for example, genomic analyses have identified 2168 pseudogenes or approximately one pseudogene for every eight functional genes (Harrison *et al.*, 2001). In humans, one pseudogene was identified for approximately every two functional genes (Harrison *et al.*, 2002). As pseudogenes generally do not confer a selective advantage, they have a low probability of being fixed in large populations (Ober, 2010).

Unless the presence of an extra amount of gene product is advantageous, it is unlikely that two genes with the same function will be stably maintained in the genome of the organism (Nowak *et al.*, 1997). In subfunctionalization (Figure 2c), both duplicated copies may become, by accumulation of mutations, partially compromised to the point at which their total capacity is reduced to the level of the single-copy ancestral gene (Force *et al.*, 1999; Stoltzfus, 1999; Lynch and Force, 2000). Subfunctionalization can occur through the modification of the regulatory elements by mutations (Force *et al.*, 1999; Hinman and Davidson, 2007) or by epigenetic silencing (Rodin and Riggs, 2003). In an evolutionary scale, one of the most important forms of subfunctionalization is the division of gene expression after duplication (Force *et al.*, 1999). For example, zebrafish ENGRAILED 1 and ENGRAILED 1-B, generated by a chromosomal segmental duplication, are a pair of transcription factors that occurred in the lineage of ray-finned fish. While ENGRAILED-1 is expressed in the pectoral appendage bud, ENGRAILED 1-B is expressed in a specific set of neurons in the hindbrain/spinal cord (Force *et al.*, 1999). In yeast, more than 40% of gene pairs exhibit significant expression divergence (Gu *et al.*, 2002). Also, the comparison of 17 fungal genomes revealed that duplicated genes rarely diverge with respect to biochemical function, but typically diverge with respect to regulatory control (Wapinski *et al.*, 2007). On the other hand, if two redundant gene copies were retained without significant functional divergence in the genome, the organism may acquire increased genetic robustness against harmful mutations (Figure 1d) (Conrad and Antonarakis, 2007).

In neofunctionalization (Figure 2b), the ancestral gene keeps its ancestral function, while the duplicated gene gains a new function under positive selection for advantageous mutations (De Grassi *et al.*, 2008). However, in many cases, rather than an entirely new function, a related function evolves after gene duplication. For example, the red and green-sensitive opsin genes of humans where the result of a gene duplication that occurred in hominoids and Old World monkeys (Yokoyama and Yokoyama, 1989). After the duplication process,

functional divergence of the two opsins resulted in a 30-nanometer difference in their maximum absorption wavelength. This difference conferred a sensitivity to a wide range of colors for humans and related primates (Zhang, 2003).

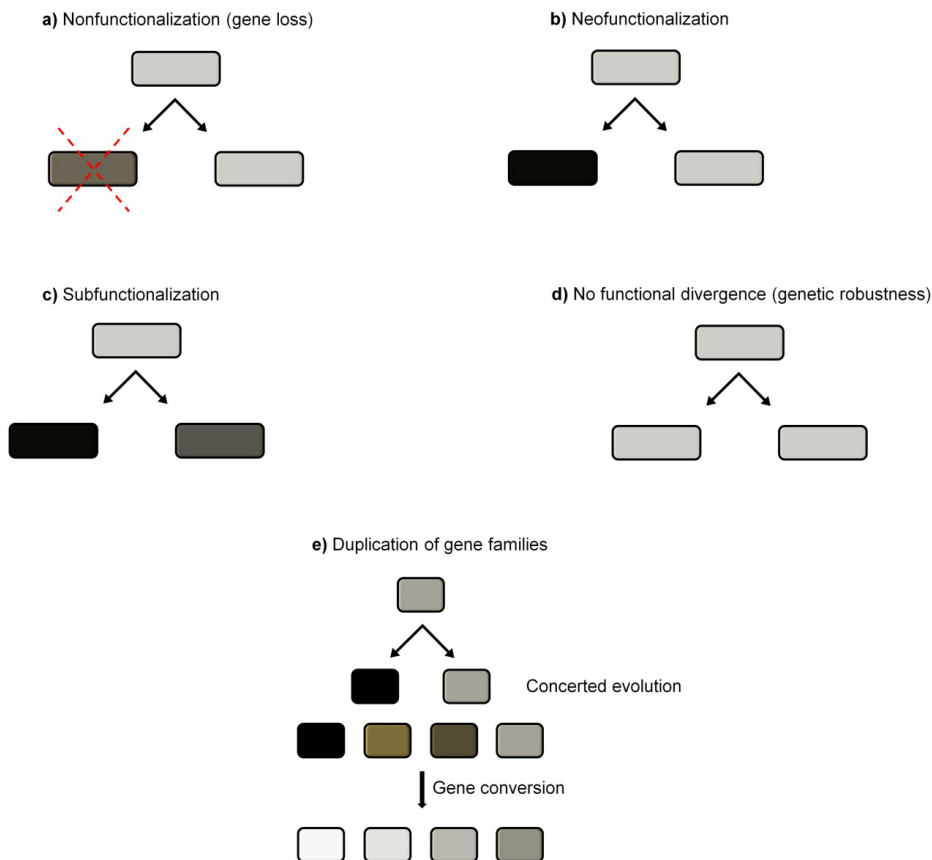


Fig. 2. Evolutionary fate of duplicated genes. Gene duplication may result in new functions via different scenarios. **(a)** nonfunctionalization; **(b)** neofunctionalization; **(c)** subfunctionalization; **(d)** genetic robustness and **(e)** gene conversion. Adapted from Conrad and Antonarakis (2007).

The fate of a gene that suffers duplication seems to be the result of diverse and, in some cases, interdependent factors (Taylor *et al.*, 2001). These variables include its functional category (Papp *et al.*, 2003; Kondrashov and Koonin, 2004; Marland *et al.*, 2004), degree of conservation (Conant and Wagner, 2002; Davis and Petrov, 2004; Jordan *et al.*, 2004; Braybrook and Harada, 2008), sensitivity to dosage effects (Kondrashov and Koonin, 2004), as well as its regulatory and architectural complexity (He and Zhang, 2005). Some observations indicate that natural selection created a preferential association of duplications with certain gene categories. For example, genes encoding proteins that interact with the environment are more frequently retained after the duplication process than genes which interact at intracellular compartments

(Li *et al.*, 2003; Marland *et al.*, 2004). In addition, genomes tend to retain duplicated genes involved in signal transduction and transcription, but to lose duplicated DNA repair genes (Blanc and Wolfe, 2004; Maere *et al.*, 2005; Paterson *et al.*, 2010).

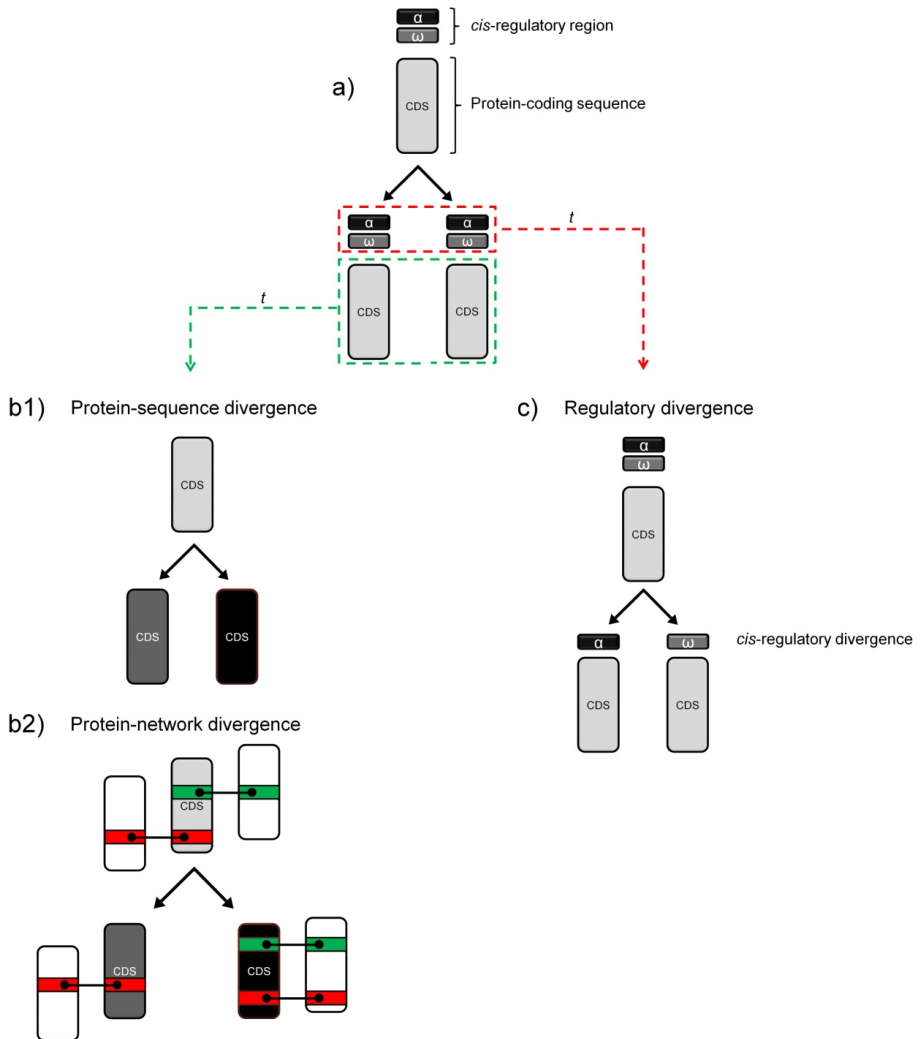


Fig. 3. Functional divergence of duplicated genes. **(a)** The independent evolution of *cis*-regulatory and the protein-coding regions. **(b1)** protein sequence divergence after duplication; **(b2)** protein network divergence with increase or loss of partners and **(c)** DNA sequence regulatory divergence after duplication (certain regulatory motifs are lost in one copy of the duplicated gene sequence).  $t$ ; evolutionary time. Adapted from Conrad and Antonarakis (2007).

It has been shown that shortly after duplication the protein-coding sequence and *cis*-regulatory regions of some duplicated genes can evolve independently (Figure 3a) (Wagner, 2000). This independent evolution can generate protein sequence divergence of duplicated genes (Figure 3b1) or protein network divergence (Figure 3b2), where the protein interaction domains (*cis*-regulatory elements) of the original sequence evolve by maintenance, gain, or loss of interacting partners. Alternatively, the divergence of *cis*-regulatory motifs in the promoter-proximal region (Figure 3c) can generate expression divergence between the duplicated genes (Conrad and Antonarakis, 2007).

## 5. Gene duplication of NF-Y in plants

While duplication of NF-Y genes is poorly understood in the plant lineage, many of the functional mechanistic details are likely conserved across plant, animal and fungal lineages. This inference comes from strong cross-kingdom conservation of functional important amino acid residues in mammalian and yeast NF-Ys (Maity and de Crombrughe, 1992; Maity *et al.*, 1992; Sinha *et al.*, 1995; Coustry *et al.*, 1996; Kim *et al.*, 1996; Sinha *et al.*, 1996; Mantovani, 1998; Romier *et al.*, 2003). CCAAT-like motifs are found in several plant promoters, and binding activity to CCAAT sequences has been identified in plant nuclear extracts (Yazawa and Kamada, 2007). Besides, at least some plant NF-YA and NF-YB subunits have been shown to complement yeast mutant strains lacking the corresponding NF-Y subunit. Additionally, several groups have demonstrated that each of the three plant NF-Y proteins can substitute their yeast counterparts in gene expression assays (Edwards *et al.*, 1998; Masiero *et al.*, 2002; Ben-Naim *et al.*, 2006; Siefers *et al.*, 2009). These observations indicate that plant NF-Y subunits might act as general transcription factors, as in mammals (Yamamoto *et al.*, 2009).

Although a complete functional plant NF-Y complex has not yet been described, the individual subunits are known to be involved in a number of important physiological processes, such as specific developmental processes and response to environmental stimuli (Lotan *et al.*, 1998; Kusnetsov *et al.*, 1999; Miyoshi *et al.*, 2003; Ben-Naim *et al.*, 2006; Combier *et al.*, 2006; Wenkel *et al.*, 2006; Cai *et al.*, 2007; Nelson *et al.*, 2007; Warpeha *et al.*, 2007; Siefers *et al.*, 2009). A well-established example is the NF-YB subunit gene called LEAFY COTYLEDON-1 (LEC1), which specifically controls embryo development, especially the maturation phase. LEC1 plays specialized roles not only because of its developmentally regulated expression but also due to its distinct molecular activity, as the *in vivo* function of LEC1 cannot be replaced by other NF-YB subunits, except for the most closely related Leafy Cotyledon 1 Like (L1L) (Kwong *et al.*, 2003; Lee *et al.*, 2003; Yamamoto *et al.*, 2009). In Arabidopsis, many NF-Y subunit genes are expressed ubiquitously, although some are differentially expressed. For example, while the AtNF-YC-4 transcript accumulates in seeds 7 days after germination, AtNF-YB-9 is only expressed in green siliques (Gusmaroli *et al.*, 2001).

Plant NF-Y function also appears to be important for responses to drought stress. Although a specific mechanism of action remains unclear, overexpression of the AtNF-YB1 subunit and its orthologue in maize (*Zea mays*), ZmNF-YB2, leads to enhanced drought resistance (Nelson *et al.*, 2007). Another study showed that overexpression of maize NF-YA5 reduced drought susceptibility, anthocyanin production and stomatal aperture, while *nf-ya5* mutants had the expected opposite phenotype in each situation (Li *et al.*, 2008). In addition, several

publications strongly suggest that NF-Y transcription factors are also involved in photoperiod-regulated flowering (Ben-Naim *et al.*, 2006; Wenkel *et al.*, 2006; Siefers *et al.*, 2009).

We adopted a high throughput comparative genomic approach to conduct a broad survey of fully sequenced genomes, including representatives of amoebozoa, yeasts, fungi, algae, mosses, plants, vertebrate and invertebrate species to identify the presence of homologous genes coding for each of the three subunits that form the NF-Y transcription factor (Table 1). NF-Y gene and protein sequences were obtained through blast searches (blastp, blastx and tblastx) against the Protein and Genome databases with the default parameters at the NCBI (National Center for Biotechnology Information - <http://www.ncbi.nlm.nih.gov>) and against completed genome projects database at the JGI (Joint Genome Institute - <http://www.jgi.doe.gov>).

The results point to a scenario where all fungi and the majority of metazoa possess single genes coding for each of the NF-Y subunits (Table 1). The metazoa exceptions include the amphioxus *Branchiostoma floridae*, the nematode *Caenorhabditis elegans* and the gastropod *Lottia gigantea*, all of each present a proportional duplication of the three subunits, possessing two genes for each subunit (Table 1).

In contrast, plants possess gene families coding for each NF-Y subunit (Table 1). For instance, in the model plant *Arabidopsis thaliana* 10 genes coding for NF-YA, 13 for NFY-B, and 13 for NF-YC were identified. Because of the heterotrimeric composition, the 36 *Arabidopsis* NF-Y subunits could theoretically combine to generate 1.690 unique transcription factors (Siefers *et al.*, 2009). This *Arabidopsis* NF-Y expansion is a general feature of the plant lineage, including monocots and eudicots. In rice (*Oryza sativa*), for example, 11 genes were identified coding for the NF-YA subunit, 10 for NF-YB and 8 for NFY-C. Four of the rice NF-YB subunits have been characterized and at least one of these genes is involved in chloroplast development (Miyoshi *et al.*, 2003; Yazawa and Kamada, 2007). Interestingly, the moss *Physcomitrella patens* and the lycophyte *Selaginella mollendorffii* possess single genes coding for NF-YA subunits whereas the other subunits are encoded by multiple genes (Table 1).

Since the evolutionary rates can be species dependent, the difference observed in the number of genes of NF-Y subunits in eukaryotic class (Table 1), especially in vascular plants, can be result of recent duplication process that contribute to the establishment of genes families coding each NF-Y subunit. However, some duplicated genes might have suffered high level of diversification what could be responsible to prevent their identification in our analyses.

Representative plants genes (monocot and eudicot) were selected to perform phylogenetic analyses of the NF-Y subunits. The phylogenetic analysis was reconstructed after protein sequence alignments using a Bayesian approach in MrBayes 3.1.2 (Ronquist and Huelsenbeck, 2003). The mixed amino acid substitution model plus gamma and invariant sites was used in two independent runs of 5,000,000 generations each with two Metropolis-coupled Monte Carlo Markov chains (MCMCMC) that were run in parallel (starting each from a random tree). Markov chains were sampled every 100 generations, and the first 25% of the trees were discarded as burn-in. The remaining ones were used to compute the majority rule consensus tree, the posterior probability of clades and branch lengths (Figure 4 to 6).

Phylo	Specie	Code	Subunit A Genes	Subunit B Genes	Subunit C Genes
<b>Metazoa</b>	<i>Homo sapiens</i>	Hsa	1	1	1
	<i>Mus musculus</i>	Mmu	1	1	1
	<i>Rattus norvegicus</i>	Rno	1	1	1
	<i>Canis familiaris</i>	Cfa	1	1	1
	<i>Monodelphis domestica</i>	Mdo	1	1	1
	<i>Gallus gallus</i>	Gga	1	1	1
	<i>Xenopus tropicalis</i>	Xtr	1	1	1
	<i>Gasterosteus aculeatus</i>	Gac	1	1	1
	<i>Oryzias latipes</i>	Ola	1	1	1
	<i>Takifugu rubripes</i>	Tru	1	1	1
	<i>Danio rerio</i>	Dre	1	1	1
	<i>Ciona savignyi</i>	Csa	1	1	1
	<i>Branchiostoma floridae</i>	Bfl	2	2	2
	<i>Strongylocentrotus purpuratus</i>	Spu	1	1	1
	<i>Drosophila melanogaster</i>	Dme	1	1	1
	<i>Anopheles gambiae</i>	Aga	1	1	1
	<i>Tribolium castaneum</i>	Tca	1	1	1
	<i>Caenorhabditis elegans</i>	Cel	2	2	2
	<i>Lottia gigantea</i>	Lgi	2	2	2
	<i>Nematostella vectensis</i>	Nve	1	1	1
<b>Fungi</b>	<i>Neurospora crassa</i>	Ncr	1	1	1
	<i>Candida tropicalis</i>	Ctr	1	1	1
	<i>Tuber melanosporum</i>	Tme	1	1	1
	<i>Pyrenophora teres</i>	Pte	1	1	1
	<i>Aspergillus nidulans</i>	Ani	1	1	1
	<i>Chaetomium globosum</i>	Cgl	1	1	1
	<i>Penicillium marneffei</i>	Pma	1	1	1
	<i>Talaromyces stipitatus</i>	Tst	1	1	1
	<i>Sordaria macrospora</i>	Sma	1	1	1
<b>Heterolobosea</b>	<i>Naegleria gruberi</i>	Ngr	1	1	1
<b>Metaphyta</b>	<i>Manihot esculenta</i>	Mes	12	15	9
	<i>Ricinus communis</i>	Rco	6	12	7
	<i>Populus trichocarpa</i>	Ptr	8	17	9
	<i>Medicago truncatula</i>	Mtr	5	10	5
	<i>Glycine max</i>	Gma	21	25	11
	<i>Cucumis sativus</i>	Csa	6	11	3
	<i>Prunus persica</i>	Ppe	6	13	6
	<i>Arabidopsis thaliana</i>	Ath	10	13	13
	<i>Carica papaya</i>	Cpa	5	9	3
	<i>Vitis vinifera</i>	Vvi	7	12	5
	<i>Sorghum bicolor</i>	Sbi	9	10	7
	<i>Zea mays</i>	Zma	10	20	14
	<i>Oryza sativa</i>	Osa	11	10	8
	<i>Brachipodium</i>	Bdi	7	13	10

distachyon					
	Selaginella mollendorffii	Smo	1	5	3
	Physcomitrella patens	Ppa	1	6	6
Heterokonta	Phaeodactylum	Ptri	1	1	1
	tricornutum				

Table 1. NF-Y genes identified in the fully eukaryotic sequenced genomes.

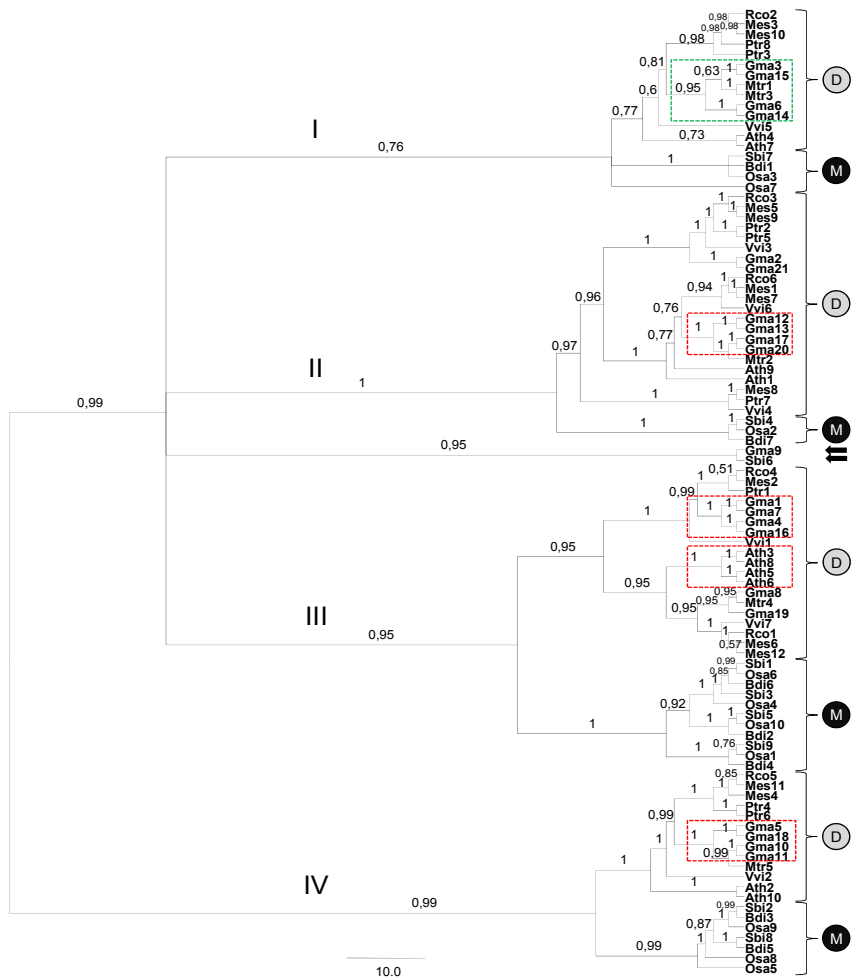


Fig. 4. Phylogenetic tree of monocot and eudicot representatives of NF-YA subunit. M: monocots; D: eudicots; Rco: *Ricinus communis*; Mes: *Manihot esculenta*; Ptr: *Populus trichocarpa*; Gma: *Glycine max*; Mtr: *Medicago truncatula*; Vvi: *Vitis vinifera*; Ath: *Arabidopsis thaliana*; Sbi: *Sorghum bicolor*; Bdi: *Brachipodyum distachyon*; Osa: *Oryza sativa*; red square: event of duplication inside the specie; green square: event of duplication inside the same plant family; black arrows: genes that possess an unresolved position in the phylogenetic tree; I to IV: independent phylogenetic gene clusters.



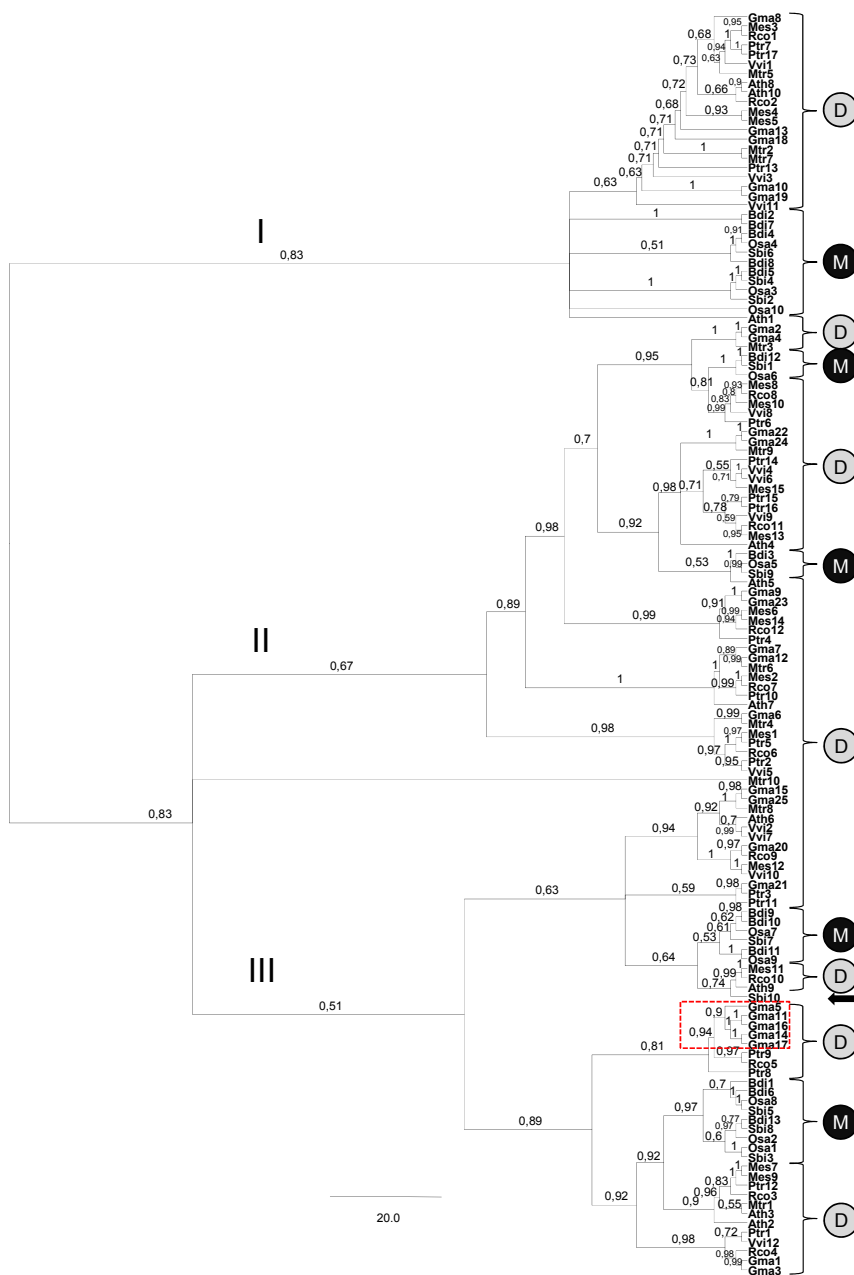


Fig. 5. Phylogenetic tree of monocot and eudicot representatives of NF-YB subunit.  
For details see legend of Figure 4.

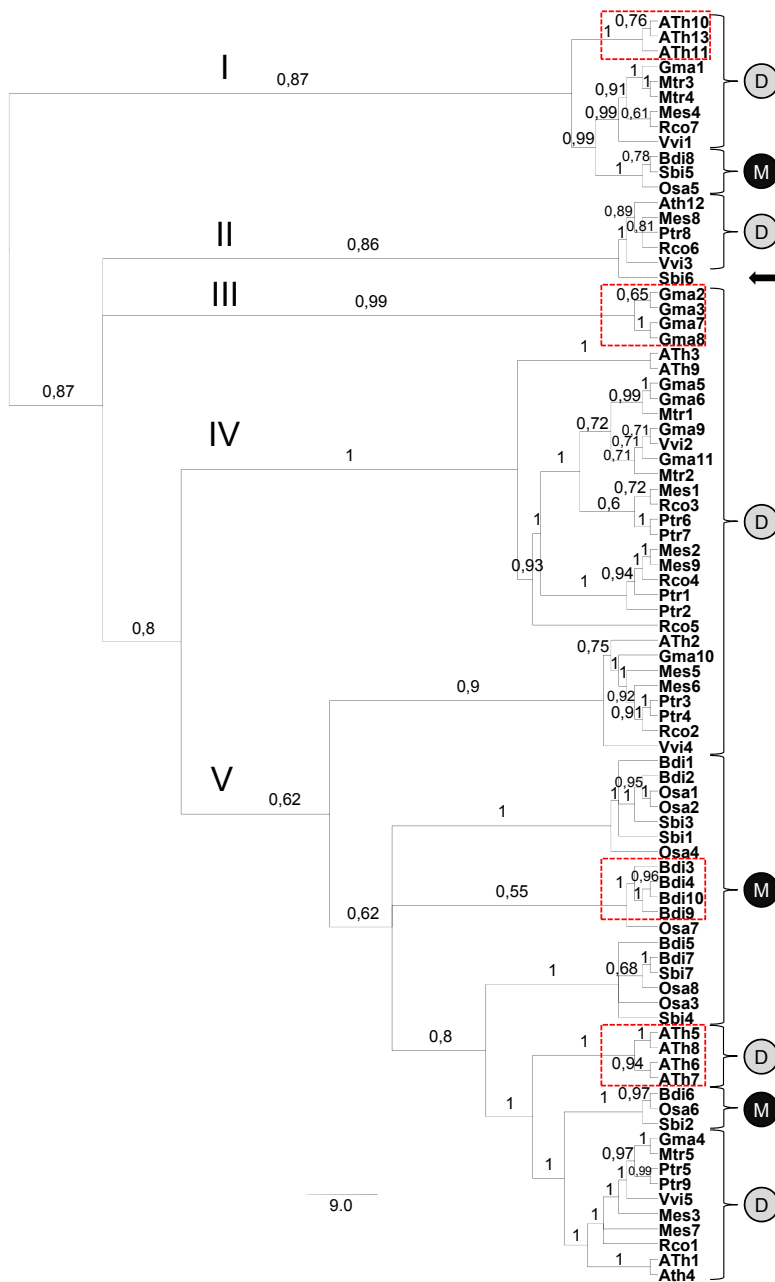


Fig. 6. Phylogenetic tree of monocot and eudicot representatives of NF-YC subunit. For details see legend of Figure 4.

Phylogenetic analysis showed that the gene diversification of all NF-Y subunits likely resulted from several duplication events along evolution and diversification of plants (Figure 4 to 6). It was possible to observe the formation of four independent highly supported clusters for the NF-YA subunit (I to IV, Figure 4), three for NF-YB (I to III, Figure 5) and five for NF-YC (I to V, Figure 6). Based on these results, we suggest that each cluster might possess an independent ancestral subunit that the duplicated members of each group originated from. However, independent duplication events have occurred in many species after the divergence of monocots and eudicots. For example, the soybean and Arabidopsis genomes have experienced a series of recent duplication events (red squares in figures 4 to 6) that could be the result of chromosome duplication or could be derived from polyploidization events (soybean is a good example of a recent polyploidization). These duplications can help us to explain the differences observed in the number of genes coding for the NF-Y subunits in plants (Table 1). Additionally, these duplications seem to be relatively recent and can provide the raw material for neofunctionalization (Figure 2b) and functional divergence of duplicated genes (Figure 3). With few exceptions (genes that possess an unresolved position in the phylogenetic tree are plotted with black arrows, Figures 4 to 6), all clusters of a specific NF-Y subunit are formed by well-defined sub-clusters of monocot and eudicot representatives (Figure 4 to 6). Events of duplication inside a specific plant family were also observed between the two fabaceae species *Glycine max* and *Medicago truncatula* (green square, Figure 4), which could indicate concerted evolution of duplicated genes between these related species (Figure 2e). This is similar to the clade-specific shifts in selective constraint following concerted duplication events observed for MADS box transcription factors in angiosperms (Shan *et al.*, 2009).

The duplication process is a prominent feature of plant genomic architecture (Figure 1). This has led many researchers to speculate that gene duplication may have played an important role in the evolution of phenotypic novelty within the plant lineage (Flagel and Wendel, 2009). As a result of pervasive and recurring small-scale duplications, which may be followed by functional divergence, many nuclear genes in plants are members of gene families and may exhibit copy number variation lineages (Blanc and Wolfe, 2004; Schlueter *et al.*, 2004), as can be observed in table 1.

Evidence for frequent gene duplication has also been observed in the evolutionary history of numerous gene families that have expanded during the diversification of the angiosperms (De Bodt *et al.*, 2005; Zahn *et al.*, 2005; Duarte *et al.*, 2010). In multigene families descended from a common ancestor, individual genes in the group exert similar functions and have similar DNA sequences (Conrad and Antonarakis, 2007). One concept, concerted evolution, applies particularly to localized and typically tandem copies of a gene. The concept posits that all genes in a given group evolve coordinately, and that homogenization is the result of gene conversion (Figure 2e) (Conrad and Antonarakis, 2007).

The emerging picture points to plant NF-Y complexes acting as essential regulatory hubs for many processes. Multiple NF-Y subunits in vascular plants may associate with each other in various combinations that regulate the expression of specific gene sets and might provide similar levels of combinatorial diversity for transcriptional fine-tuning (Siefers *et al.*, 2009). The amplification observed in the plant lineage (Table 1) raises the possibility that new and divergent functions of heterotrimeric complexes have evolved in plants (Nelson *et al.*, 2007) indicating a more complex regulatory role for the various NF-Y proteins in plants than in other organisms (Stephenson *et al.*, 2007).

The existence of multiple genes for each subunit in the plant genome indicates that the specificity of subunit interaction may be determined by preferential protein-protein interaction, tissue or cell-specific expression of each gene or a combination of both (Yazawa and Kamada, 2007). The large number of possible combinations has hindered the analysis of plant NF-Y complexes and suggests that they might act in a more intricate system than in vertebrates and yeast, which have only one gene that encodes each HAP subunit (Yazawa and Kamada, 2007). Additionally, the multiple copies for each NF-Y subunit raises a question if a specific NF-Y subunit interacts with any other two NF-Y subunits or if the NF-Y subunit interacts with only specific member(s) of the other two subunits (Thirumurugan *et al.*, 2008).

Although the presence of many genes encoding NF-Y subunits suggests a high degree of genetic redundancy in plants, the analysis of mutants in single NF-Y genes in *Arabidopsis* has been associated with defects in development and enhanced stress sensitivity, suggesting a specialized function for each member (Lotan *et al.*, 1998; Kwong *et al.*, 2003; Lee *et al.*, 2003; Zanetti *et al.*, 2010). This could indicate that duplicated genes have passed through a neofunctionalization process (Figure 2b).

Some proteins may require several key substitutions before acquiring a new function, while others may be more mutationally labile. An example includes the terpene synthase gene family in Norway spruce (*Picea abies*). These genes appear to have undergone repeated rounds of neofunctionalization (Figure 2b) (Keeling *et al.*, 2008) and a small number of key amino acid substitutions among paralogs was sufficient to alter the substrate specificity and terpenoid product profiles (Flagel and Wendel, 2009). Another example of neofunctionalization (Figure 2b) in plants is observed in *Arabidopsis*, where a specific amino acid residue identified in LEC1 and LEC1-LIKE (L1L) is responsible for differentiating their functions (seed development) from those of other NF-YB members (Kwong *et al.*, 2003; Lee *et al.*, 2003; Yamamoto *et al.*, 2009). In addition, the analysis of amino acid substitution rates in plants has been appointed for the asymmetric evolution of certain duplicates of NF-YB and NF-YC subunits, which appears to be coupled with the asymmetric divergence in gene function (Yang *et al.*, 2005; Yamamoto *et al.*, 2009).

With respect to expression patterns, the *Arabidopsis* NF-Y gene family presents some members that are ubiquitously expressed and others that are tissue specific or induced only after the switch to reproductive growth in flowers and siliques (Gusmaroli *et al.*, 2001; 2002; Yazawa and Kamada, 2007). The difference observed in the expression pattern of these genes could represent an example of *cis*-regulatory divergence (Figure 3c), where the *cis*-element of gene evolves independently from the other members of gene family, and becomes regulated by different stimuli and/or *trans*-activators.

Because genes that harbor NF-Y binding domains include genes that are constitutive, inducible, and cell-cycle-dependent, the regulation of the expression of these genes cannot be exclusively due to NF-Y binding to DNA. In this scenario, the interaction with other transcription factors, either functionally or physically, will contribute to the NF-Y action (Matuoka and Chen, 2002). In addition, the independent evolution of protein-binding domains present in duplicated gene architecture can contribute to protein network divergence (Figure 3b2), increasing the numbers of possible interacting partners of NF-Y genes.

When compared with other forms of mutation, a notable feature of duplication is that it creates genetic redundancy. This redundancy fosters evolutionary innovation, creating the opportunity for duplicates to explore new evolutionary terrain (Flagel and Wendel, 2009).

The most important contribution of gene duplication to evolution is to supply new genetic material for mutation, drift and selection to act upon. This leads to the creation of new genes and new gene functions (Hurley *et al.*, 2005; Woollard, 2005; Schmidt and Davies, 2007), two important factors in the origin of genomic and organismal complexity (Gu *et al.*, 2002; Taylor and Raes, 2004; Sterck *et al.*, 2007).

The plasticity of a genome or species in adapting to environmental changes would be severely limited without gene duplication, because no more than two variants (alleles) exist at any locus within a diploid individual. A good example is the dozens of duplicated immunoglobulin genes that constitute the vertebrate adaptive immune system. It seems difficult to imagine how this system could have acquired this high complexity level without gene duplication (Zhang, 2003).

Plant gene families are largely conserved even over evolutionary time scales that encompass the diversification of all angiosperms and nonflowering plants (Rensing *et al.*, 2008). This property of plant genomes indicates that plants have not created new gene families, but have been endowed with a basic genetic toolkit of ancient origin. Despite the evolutionary conservation of gene families, lineage-specific fluctuations in gene family size are frequently observed among taxa (Velasco *et al.*, 2007; Ming *et al.*, 2008; Rensing *et al.*, 2008), which suggests that the diversity and lineage-specific phenotypic variation observed in land plants may not be explained by an equally diverse set of entirely novel genes. Indeed, much of plant diversity may have arisen from the duplication and adaptive specialization processes of pre-existing genes (neofunctionalization and subfunctionalization, Figure 2b and c, respectively). This perspective assigns gene duplication a central role in plant diversification, being a key process that generates the raw material necessary for adaptive evolution (Flagel and Wendel, 2009).

## 6. Conclusions

Whereas various classes of structural and metabolic genes preferentially return to a single copy state following whole-genome duplication (Paterson *et al.*, 2010), transcription factors tend to be preferentially retained among the duplicated genes in *A. thaliana* (Flagel and Wendel, 2009). Our findings support the hypotheses that this preference seems to be true for all plant species, based on the number of genes identified for each NF-Y subunit. Certainly, further studies encompassing functional assays are required to ascertain the role of these genes in plant metabolism.

The number of interacting partners in a molecular network (connectivity) of a particular gene also influences the probability of duplication gene retention (Flagel and Wendel, 2009). In this scenario, the high number of genes coding for the three subunits of NF-Y transcription factor in higher plants leads to numerous interaction possibilities among different genes of each subunit and among these genes and other transcription factors what could contribute to gene retention of the NF-Y transcription factor family in plants.

## 7. Acknowledgement

This work was supported by a CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), FAPERGS (Fundação de Amparo a Pesquisa do Estado do Rio Grande do Sul), FINEP (Financiadora de Projetos) and MCT (Ministério de Ciência e Tecnologia). A. Cagliari

received a Ph.D. fellowship from CNPq. and G. Loss a Ph.D. fellowship from CAPES. A. Turchetto-Zolet and F. Maraschin have a PNPd-CAPES fellowship and M. Margis-Pinheiro and R. Margis are recipients of CNPq research fellowships number 308708/2006-7 and 303967/2008-0, respectively.

## 8. References

- Arents, G. and Moudrianakis, E.N. (1995) The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. *Proc Natl Acad Sci U S A*, 92, 11170-11174.
- Baxeavanis, A.D., Arents, G., Moudrianakis, E.N. and Landsman, D. (1995) A variety of DNA-binding and multimeric proteins contain the histone fold motif. *Nucleic Acids Res*, 23, 2685-2691.
- Beisswanger, S. and Stephan, W. (2008) Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes in *Drosophila*. *Proc Natl Acad Sci U S A*, 105, 5447-5452.
- Ben-Naim, O., Eshed, R., Parnis, A., Teper-Bamnolker, P., Shalit, A., Coupland, G., Samach, A. and Lifschitz, E. (2006) The CCAAT binding factor can mediate interactions between CONSTANS-like proteins and DNA. *Plant J*, 46, 462-476.
- Berry, M., Grosveld, F. and Dillon, N. (1992) A single point mutation is the cause of the Greek form of hereditary persistence of fetal haemoglobin. *Nature*, 358, 499-502.
- Blanc, G. and Wolfe, K.H. (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, 16, 1667-1678.
- Bowers, J.E., Chapman, B.A., Rong, J. and Paterson, A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422, 433-438.
- Braybrook, S.A. and Harada, J.J. (2008) LECs go crazy in embryo development. *Trends Plant Sci*, 13, 624-630.
- Bucher, P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol*, 212, 563-578.
- Cai, X., Ballif, J., Endo, S., Davis, E., Liang, M., Chen, D., DeWald, D., Kreps, J., Zhu, T. and Wu, Y. (2007) A putative CCAAT-binding transcription factor is a regulator of flowering timing in *Arabidopsis*. *Plant Physiol*, 145, 98-105.
- Caretti, G., Motta, M.C. and Mantovani, R. (1999) NF-Y associates with H3-H4 tetramers and octamers by multiple mechanisms. *Mol Cell Biol*, 19, 8591-8603.
- Ceribelli, M., Dolfini, D., Merico, D., Gatta, R., Vigano, A.M., Pavesi, G. and Mantovani, R. (2008) The histone-like NF-Y is a bifunctional transcription factor. *Mol Cell Biol*, 28, 2047-2058.
- Chodosh, L.A., Baldwin, A.S., Carthew, R.W. and Sharp, P.A. (1988) Human CCAAT-binding proteins have heterologous subunits. *Cell*, 53, 11-24.
- Combiér, J.P., Frugier, F., de Billy, F., Boualem, A., El-Yahyaoui, F., Moreau, S., Vernie, T., Ott, T., Gamas, P., Crespi, M. and Niebel, A. (2006) MtHAP2-1 is a key transcriptional regulator of symbiotic nodule development regulated by microRNA169 in *Medicago truncatula*. *Genes Dev*, 20, 3084-3088.

- Conant, G.C. and Wagner, A. (2002) GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res*, 30, 3378-3386.
- Conrad, B. and Antonarakis, S.E. (2007) Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev Genomics Hum Genet*, 8, 17-35.
- Coustry, F., Maity, S.N., Sinha, S. and de Crombrughe, B. (1996) The transcriptional activity of the CCAAT-binding factor CBF is mediated by two distinct activation domains, one in the CBF-B subunit and the other in the CBF-C subunit. *J Biol Chem*, 271, 14485-14491.
- Coustry, F., Hu, Q., de Crombrughe, B. and Maity, S.N. (2001) CBF/NF-Y functions both in nucleosomal disruption and transcription activation of the chromatin-assembled topoisomerase II $\alpha$  promoter. Transcription activation by CBF/NF-Y in chromatin is dependent on the promoter structure. *J Biol Chem*, 276, 40621-40630.
- Currie, R.A. (1998) NF-Y is associated with the histone acetyltransferases GCN5 and P/CAF. *J Biol Chem*, 273, 1430-1434.
- Dang, V.D., Bohn, C., Bolotin-Fukuhara, M. and Daignan-Fornier, B. (1996) The CCAAT box-binding factor stimulates ammonium assimilation in *Saccharomyces cerevisiae*, defining a new cross-pathway regulation between nitrogen and carbon metabolisms. *J Bacteriol*, 178, 1842-1849.
- Davis, J.C. and Petrov, D.A. (2004) Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol*, 2, E55.
- De Bodt, S., Maere, S. and Van de Peer, Y. (2005) Genome duplication and the origin of angiosperms. *Trends Ecol Evol*, 20, 591-597.
- De Grassi, A., Lanave, C. and Saccone, C. (2008) Genome duplication and gene-family evolution: the case of three OXPHOS gene families. *Gene*, 421, 1-6.
- de Silvio, A., Imbriano, C. and Mantovani, R. (1999) Dissection of the NF-Y transcriptional activation potential. *Nucleic Acids Res*, 27, 2578-2584.
- Dorn, A., Bollekens, J., Staub, A., Benoist, C. and Mathis, D. (1987a) A multiplicity of CCAAT box-binding proteins. *Cell*, 50, 863-872.
- Dorn, A., Durand, B., Marfing, C., Le Meur, M., Benoist, C. and Mathis, D. (1987b) Conserved major histocompatibility complex class II boxes--X and Y--are transcriptional control elements and specifically bind nuclear proteins. *Proc Natl Acad Sci U S A*, 84, 6249-6253.
- Duarte, J.M., Wall, P.K., Edger, P.P., Landherr, L.L., Ma, H., Pires, J.C., Leebens-Mack, J. and dePamphilis, C.W. (2010) Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evol Biol*, 10, 61.
- Dynan, W.S. and Tjian, R. (1985) Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. *Nature*, 316, 774-778.
- Edwards, D., Murray, J.A. and Smith, A.G. (1998) Multiple genes encoding the conserved CCAAT-box transcription factor complex are expressed in *Arabidopsis*. *Plant Physiol*, 117, 1015-1022.
- Flagel, L.E. and Wendel, J.F. (2009) Gene duplication and evolutionary novelty in plants. *New Phytol*, 183, 557-564.

- Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L. and Postlethwait, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151, 1531-1545.
- Forsburg, S.L. and Guarente, L. (1989) Identification and characterization of HAP4: a third component of the CCAAT-bound HAP2/HAP3 heteromer. *Genes Dev*, 3, 1166-1178.
- Gallagher, C.E., Matthews, P.D., Li, F. and Wurtzel, E.T. (2004) Gene duplication in the carotenoid biosynthetic pathway preceded evolution of the grasses. *Plant Physiol*, 135, 1776-1783.
- Gu, Z., Nicolae, D., Lu, H.H. and Li, W.H. (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet*, 18, 609-613.
- Gusmaroli, G., Tonelli, C. and Mantovani, R. (2001) Regulation of the CCAAT-Binding NF-Y subunits in *Arabidopsis thaliana*. *Gene*, 264, 173-185.
- Gusmaroli, G., Tonelli, C. and Mantovani, R. (2002) Regulation of novel members of the *Arabidopsis thaliana* CCAAT-binding nuclear factor Y subunits. *Gene*, 283, 41-48.
- Harrison, P.M., Echols, N. and Gerstein, M.B. (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucleic Acids Res*, 29, 818-830.
- Harrison, P.M., Hegyi, H., Balasubramanian, S., Luscombe, N.M., Bertone, P., Echols, N., Johnson, T. and Gerstein, M. (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res*, 12, 272-280.
- He, X.L. and Zhang, J.Z. (2005) Gene complexity and gene duplicability. *Current Biology*, 15, 1016-1021.
- Hinman, V.F. and Davidson, E.H. (2007) Evolutionary plasticity of developmental gene regulatory network architecture. *Proc Natl Acad Sci U S A*, 104, 19404-19409.
- Holland, P.W., Garcia-Fernandez, J., Williams, N.A. and Sidow, A. (1994) Gene duplications and the origins of vertebrate development. *Dev Suppl*, 125-133.
- Huang, H.S., Chen, C.J. and Chang, W.C. (1999) The CCAAT-box binding factor NF-Y is required for the expression of phospholipid hydroperoxide glutathione peroxidase in human epidermoid carcinoma A431 cells. *FEBS Lett*, 455, 111-116.
- Hurley, I., Hale, M.E. and Prince, V.E. (2005) Duplication events and the evolution of segmental identity. *Evol Dev*, 7, 556-567.
- Jordan, I.K., Wolf, Y.I. and Koonin, E.V. (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol Biol*, 4, 22.
- Keeling, C.I., Weisshaar, S., Lin, R.P. and Bohlmann, J. (2008) Functional plasticity of paralogous diterpene synthases involved in conifer defense. *Proc Natl Acad Sci U S A*, 105, 1085-1090.
- Kim, I.S., Sinha, S., de Crombrughe, B. and Maity, S.N. (1996) Determination of functional domains in the C subunit of the CCAAT-binding factor (CBF) necessary for formation of a CBF-DNA complex: CBF-B interacts simultaneously with both the CBF-A and CBF-C subunits to form a heterotrimeric CBF molecule. *Mol Cell Biol*, 16, 4003-4013.
- Kimura, M. (1989) The neutral theory of molecular evolution and the world view of the neutralists. *Genome*, 31, 24-31.



- Kondrashov, F.A. and Koonin, E.V. (2004) A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet*, 20, 287-290.
- Kusnetsov, V., Landsberger, M., Meurer, J. and Oelmuller, R. (1999) The assembly of the CAAT-box binding complex at a photosynthesis gene promoter is regulated by light, cytokinin, and the stage of the plastids. *J Biol Chem*, 274, 36009-36014.
- Kwong, R.W., Bui, A.Q., Lee, H., Kwong, L.W., Fischer, R.L., Goldberg, R.B. and Harada, J.J. (2003) LEAFY COTYLEDON1-LIKE defines a class of regulators essential for embryo development. *Plant Cell*, 15, 5-18.
- Lee, H., Fischer, R.L., Goldberg, R.B. and Harada, J.J. (2003) Arabidopsis LEAFY COTYLEDON1 represents a functionally specialized subunit of the CCAAT binding transcription factor. *Proc Natl Acad Sci U S A*, 100, 2152-2156.
- Li, W.H., Gu, Z., Cavalcanti, A.R. and Nekrutenko, A. (2003) Detection of gene duplications and block duplications in eukaryotic genomes. *J Struct Funct Genomics*, 3, 27-34.
- Li, W.X., Oono, Y., Zhu, J., He, X.J., Wu, J.M., Iida, K., Lu, X.Y., Cui, X., Jin, H. and Zhu, J.K. (2008) The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance. *Plant Cell*, 20, 2238-2251.
- Li, X.Y., Mantovani, R., Hooft van Huijsduijnen, R., Andre, I., Benoist, C. and Mathis, D. (1992) Evolutionary variation of the CCAAT-binding transcription factor NF-Y. *Nucleic Acids Res*, 20, 1087-1091.
- Liberati, C., di Silvio, A., Ottolenghi, S. and Mantovani, R. (1999) NF-Y binding to twin CCAAT boxes: role of Q-rich domains and histone fold helices. *J Mol Biol*, 285, 1441-1455.
- Lotan, T., Ohto, M., Yee, K.M., West, M.A., Lo, R., Kwong, R.W., Yamagishi, K., Fischer, R.L., Goldberg, R.B. and Harada, J.J. (1998) Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. *Cell*, 93, 1195-1205.
- Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F. and Richmond, T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389, 251-260.
- Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, 290, 1151-1155.
- Lynch, M. and Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154, 459-473.
- Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M. and Van de Peer, Y. (2005) Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A*, 102, 5454-5459.
- Maity, S.N. and de Crombrughe, B. (1992) Biochemical analysis of the B subunit of the heteromeric CCAAT-binding factor. A DNA-binding domain and a subunit interaction domain are specified by two separate segments. *J Biol Chem*, 267, 8286-8292.
- Maity, S.N., Sinha, S., Ruteshouser, E.C. and de Crombrughe, B. (1992) Three different polypeptides are necessary for DNA binding of the mammalian heteromeric CCAAT binding factor. *J Biol Chem*, 267, 16574-16580.
- Maity, S.N. and De Crombrughe, B. (1996) Purification, characterization, and role of CCAAT-binding factor in transcription. *Methods Enzymol*, 273, 217-232.

- Maity, S.N. and de Crombrughe, B. (1998) Role of the CCAAT-binding protein CBF/NF-Y in transcription. *Trends Biochem Sci*, 23, 174-178.
- Mantovani, R. (1998) A survey of 178 NF-Y binding CCAAT boxes. *Nucleic Acids Res*, 26, 1135-1143.
- Mantovani, R. (1999) The molecular biology of the CCAAT-binding factor NF-Y. *Gene*, 239, 15-27.
- Marland, E., Prachumwat, A., Maltsev, N., Gu, Z. and Li, W.H. (2004) Higher gene duplicabilities for metabolic proteins than for nonmetabolic proteins in yeast and *E. coli*. *J Mol Evol*, 59, 806-814.
- Masiero, S., Imbriano, C., Ravasio, F., Favaro, R., Pelucchi, N., Gorla, M.S., Mantovani, R., Colombo, L. and Kater, M.M. (2002) Ternary complex formation between MADS-box transcription factors and the histone fold protein NF-YB. *J Biol Chem*, 277, 26429-26435.
- Matuoka, K. and Yu Chen, K. (1999) Nuclear factor Y (NF-Y) and cellular senescence. *Exp Cell Res*, 253, 365-371.
- Matuoka, K. and Chen, K.Y. (2002) Transcriptional regulation of cellular ageing by the CCAAT box-binding factor CBF/NF-Y. *Ageing Res Rev*, 1, 639-651.
- McNabb, D.S., Xing, Y. and Guarente, L. (1995) Cloning of yeast HAP5: a novel subunit of a heterotrimeric complex required for CCAAT binding. *Genes Dev*, 9, 47-58.
- Miyoshi, K., Ito, Y., Serizawa, A. and Kurata, N. (2003) OsHAP3 genes regulate chloroplast biogenesis in rice. *Plant J*, 36, 532-540.
- Muro, A.F., Bernath, V.A. and Kornblihtt, A.R. (1992) Interaction of the -170 cyclic AMP response element with the adjacent CCAAT box in the human fibronectin gene promoter. *J Biol Chem*, 267, 12767-12774.
- Myers, R.M., Tilly, K. and Maniatis, T. (1986) Fine structure genetic analysis of a beta-globin promoter. *Science*, 232, 613-618.
- Nelson, D.E., Repetti, P.P., Adams, T.R., Creelman, R.A., Wu, J., Warner, D.C., Anstrom, D.C., Bensen, R.J., Castiglioni, P.P., Donnarummo, M.G., Hinchey, B.S., Kumimoto, R.W., Maszle, D.R., Canales, R.D., Krolkowski, K.A., Dotson, S.B., Gutterson, N., Ratcliffe, O.J. and Heard, J.E. (2007) Plant nuclear factor Y (NF-Y) B subunits confer drought tolerance and lead to improved corn yields on water-limited acres. *Proc Natl Acad Sci U S A*, 104, 16450-16455.
- Nicole, M.C., Hamel, L.P., Morency, M.J., Beaudoin, N., Ellis, B.E. and Seguin, A. (2006) MAP-ping genomic organization and organ-specific expression profiles of poplar MAP kinases and MAP kinase kinases. *BMC Genomics*, 7, 223.
- Nowak, M.A., Boerlijst, M.C., Cooke, J. and Smith, J.M. (1997) Evolution of genetic redundancy. *Nature*, 388, 167-171.
- Ober, D. (2010) Gene duplications and the time thereafter - examples from plant secondary metabolism. *Plant Biol (Stuttg)*, 12, 570-577.
- Olesen, J.T. and Guarente, L. (1990) The HAP2 subunit of yeast CCAAT transcriptional activator contains adjacent domains for subunit association and DNA recognition: model for the HAP2/3/4 complex. *Genes Dev*, 4, 1714-1729.
- Papp, B., Pal, C. and Hurst, L.D. (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424, 194-197.

- Paterson, A.H., Freeling, M., Tang, H. and Wang, X. (2010) Insights from the comparison of plant genome sequences. *Annu Rev Plant Biol*, 61, 349-372.
- Peng, Y. and Jahroudi, N. (2002) The NFY transcription factor functions as a repressor and activator of the von Willebrand factor promoter. *Blood*, 99, 2408-2417.
- Peng, Y. and Jahroudi, N. (2003) The NFY transcription factor inhibits von Willebrand factor promoter activation in non-endothelial cells through recruitment of histone deacetylases. *J Biol Chem*, 278, 8385-8394.
- Remenyi, A., Scholer, H.R. and Wilmanns, M. (2004) Combinatorial control of gene expression. *Nat Struct Mol Biol*, 11, 812-815.
- Riechmann, J.L. and Ratcliffe, O.J. (2000) A genomic perspective on plant transcription factors. *Curr Opin Plant Biol*, 3, 423-434.
- Rieping, M. and Schoffl, F. (1992) Synergistic effect of upstream sequences, CCAAT box elements, and HSE sequences for enhanced expression of chimaeric heat shock genes in transgenic tobacco. *Mol Gen Genet*, 231, 226-232.
- Rodin, S.N. and Riggs, A.D. (2003) Epigenetic silencing may aid evolution by gene duplication. *J Mol Evol*, 56, 718-729.
- Romier, C., Cocchiarella, F., Mantovani, R. and Moras, D. (2003) The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y. *J Biol Chem*, 278, 1336-1345.
- Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19, 1572-1574.
- Roy, B. and Lee, A.S. (1995) Transduction of calcium stress through interaction of the human transcription factor CBF with the proximal CCAAT regulatory element of the *grp78/BiP* promoter. *Mol Cell Biol*, 15, 2263-2274.
- Sandman, K., Krzycki, J.A., Dobrinski, B., Lurz, R. and Reeve, J.N. (1990) HMf, a DNA-binding protein isolated from the hyperthermophilic archaeon *Methanothermus fervidus*, is most closely related to histones. *Proc Natl Acad Sci U S A*, 87, 5788-5791.
- Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J. and Shoemaker, R.C. (2004) Mining EST databases to resolve evolutionary events in major crop species. *Genome*, 47, 868-876.
- Schmidt, E.E. and Davies, C.J. (2007) The origins of polypeptide domains. *Bioessays*, 29, 262-270.
- Serra, E., Zemzoumi, K., di Silvio, A., Mantovani, R., Lardans, V. and Dissous, C. (1998) Conservation and divergence of NF-Y transcriptional activation function. *Nucleic Acids Res*, 26, 3800-3805.
- Shan, H., Zahn, L., Guindon, S., Wall, P.K., Kong, H., Ma, H., DePamphilis, C.W. and Leebens-Mack, J. (2009) Evolution of plant MADS box transcription factors: evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications. *Mol Biol Evol*, 26, 2229-2244.
- Shi, Q., Gross, K.W. and Sigmund, C.D. (2001) NF-Y antagonizes renin enhancer function by blocking stimulatory transcription factors. *Hypertension*, 38, 332-336.
- Sidow, A. (1996) Gen(om)e duplications in the evolution of early vertebrates. *Curr Opin Genet Dev*, 6, 715-722.

- Siefers, N., Dang, K.K., Kumimoto, R.W., Bynum, W.E.t., Tayrose, G. and Holt, B.F., 3rd (2009) Tissue-specific expression patterns of Arabidopsis NF-Y transcription factors suggest potential for extensive combinatorial complexity. *Plant Physiol*, 149, 625-641.
- Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M. and Van de Peer, Y. (2002) The hidden duplication past of Arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 99, 13627-13632.
- Singh, K.B. (1998) Transcriptional regulation in plants: the importance of combinatorial control. *Plant Physiol*, 118, 1111-1120.
- Sinha, S., Maity, S.N., Lu, J. and de Crombrughe, B. (1995) Recombinant rat CBF-C, the third subunit of CBF/NFY, allows formation of a protein-DNA complex with CBF-A and CBF-B and with yeast HAP2 and HAP3. *Proc Natl Acad Sci U S A*, 92, 1624-1628.
- Sinha, S., Kim, I.S., Sohn, K.Y., de Crombrughe, B. and Maity, S.N. (1996) Three classes of mutations in the A subunit of the CCAAT-binding factor CBF delineate functional domains involved in the three-step assembly of the CBF-DNA complex. *Mol Cell Biol*, 16, 328-337.
- Soltis, P.S. (2005) Ancient and recent polyploidy in angiosperms. *New Phytol*, 166, 5-8.
- Steidl, S., Papagiannopoulos, P., Litzka, O., Andrianopoulos, A., Davis, M.A., Brakhage, A.A. and Hynes, M.J. (1999) AnCF, the CCAAT binding complex of *Aspergillus nidulans*, contains products of the hapB, hapC, and hapE genes and is required for activation by the pathway-specific regulatory gene amdR. *Mol Cell Biol*, 19, 99-106.
- Stephenson, T.J., McIntyre, C.L., Collet, C. and Xue, G.P. (2007) Genome-wide identification and expression analysis of the NF-Y family of transcription factors in *Triticum aestivum*. *Plant Mol Biol*, 65, 77-92.
- Sterck, L., Rombauts, S., Vandepoele, K., Rouze, P. and Van de Peer, Y. (2007) How many genes are there in plants (... and why are they there)? *Curr Opin Plant Biol*, 10, 199-203.
- Stoltzfus, A. (1999) On the possibility of constructive neutral evolution. *J Mol Evol*, 49, 169-181.
- Struhl, K. and Moqtaderi, Z. (1998) The TAFs in the HAT. *Cell*, 94, 1-4.
- Tasanen, K., Oikarinen, J., Kivirikko, K.I. and Pihlajaniemi, T. (1992) Promoter of the gene for the multifunctional protein disulfide isomerase polypeptide. Functional significance of the six CCAAT boxes and other promoter elements. *J Biol Chem*, 267, 11513-11519.
- Taylor, J.S., Van de Peer, Y. and Meyer, A. (2001) Genome duplication, divergent resolution and speciation. *Trends Genet*, 17, 299-301.
- Taylor, J.S. and Raes, J. (2004) Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet*, 38, 615-643.
- Testa, A., Donati, G., Yan, P., Romani, F., Huang, T.H., Vigano, M.A. and Mantovani, R. (2005) Chromatin immunoprecipitation (ChIP) on chip experiments uncover a widespread distribution of NF-Y binding CCAAT sites outside of core promoters. *J Biol Chem*, 280, 13606-13615.

- Thirumurugan, T., Ito, Y., Kubo, T., Serizawa, A. and Kurata, N. (2008) Identification, characterization and interaction of HAP family genes in rice. *Mol Genet Genomics*, 279, 279-289.
- Vision, T.J., Brown, D.G. and Tanksley, S.D. (2000) The origins of genomic duplications in *Arabidopsis*. *Science*, 290, 2114-2117.
- Wagner, A. (2000) Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proc Natl Acad Sci U S A*, 97, 6579-6584.
- Wapinski, I., Pfeffer, A., Friedman, N. and Regev, A. (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449, 54-61.
- Warpeha, K.M., Upadhyay, S., Yeh, J., Adamiak, J., Hawkins, S.I., Lapik, Y.R., Anderson, M.B. and Kaufman, L.S. (2007) The GCR1, GPA1, PRN1, NF-Y signal chain mediates both blue light and abscisic acid responses in *Arabidopsis*. *Plant Physiol*, 143, 1590-1600.
- Wenkel, S., Turck, F., Singer, K., Gissot, L., Le Gourrierc, J., Samach, A. and Coupland, G. (2006) CONSTANS and the CCAAT box binding complex share a functionally important domain and interact to regulate flowering of *Arabidopsis*. *Plant Cell*, 18, 2971-2984.
- Wolberger, C. (1998) Combinatorial transcription factors. *Curr Opin Genet Dev*, 8, 552-559.
- Woollard, A. (2005) Gene duplications and genetic redundancy in *C. elegans*. *WormBook*, 1-6.
- Xiong, A.S., Peng, R.H., Zhuang, J., Gao, F., Zhu, B., Fu, X.Y., Xue, Y., Jin, X.F., Tian, Y.S., Zhao, W. and Yao, Q.H. (2009) Gene duplication, transfer, and evolution in the chloroplast genome. *Biotechnol Adv*, 27, 340-347.
- Yamamoto, A., Kagaya, Y., Toyoshima, R., Kagaya, M., Takeda, S. and Hattori, T. (2009) *Arabidopsis* NF-YB subunits LEC1 and LEC1-LIKE activate transcription by interacting with seed-specific ABRE-binding factors. *Plant J*, 58, 843-856.
- Yang, J., Xie, Z. and Glover, B.J. (2005) Asymmetric evolution of duplicate genes encoding the CCAAT-binding factor NF-Y in plant genomes. *New Phytol*, 165, 623-631.
- Yang, X., Tuskan, G.A. and Cheng, M.Z. (2006) Divergence of the Dof gene families in poplar, *Arabidopsis*, and rice suggests multiple modes of gene evolution after duplication. *Plant Physiol*, 142, 820-830.
- Yazawa, K. and Kamada, H. (2007) Identification and characterization of carrot HAP factors that form a complex with the embryo-specific transcription factor C-LEC1. *J Exp Bot*, 58, 3819-3828.
- Yokoyama, S. and Yokoyama, R. (1989) Molecular evolution of human visual pigment genes. *Mol Biol Evol*, 6, 186-197.
- Zahn, L.M., Leebens-Mack, J., DePamphilis, C.W., Ma, H. and Theissen, G. (2005) To B or Not to B a flower: the role of DEFICIENS and GLOBOSA orthologs in the evolution of the angiosperms. *J Hered*, 96, 225-240.
- Zanetti, M.E., Blanco, F.A., Beker, M.P., Battaglia, M. and Aguilar, O.M. (2010) A C subunit of the plant nuclear factor NF-Y required for rhizobial infection and nodule development affects partner selection in the common bean-Rhizobium *etli* symbiosis. *Plant Cell*, 22, 4142-4157.

- Zhang, J.Z. (2003) Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18, 292-298.
- Zwicker, J. and Muller, R. (1997) Cell-cycle regulation of gene expression by transcriptional repression. *Trends Genet*, 13, 3-6.

## **Part 3**

### **Examining Bundles of Genes**





# L- Myo-Inositol 1-Phosphate Synthase (MIPS) in Chickpea: Gene Duplication and Functional Divergence

Manoj Majee and Harmmeet Kaur

*National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi  
India*

## 1. Introduction

Gene duplication is one of the key driving forces in the evolution of genes and important features of genomic architecture of living organisms including plants. Moreover, much of the plant diversity may have arisen largely due to duplication, followed by divergence and adaptive specialization of the pre existing genes (Ohno,1970; Zhang, 2003; Flagel & Wendel,2009). Current impetus on genomic sequence data provides substantial evidence for the profusion of duplicated genes in all organisms surveyed. Functional divergence after gene duplication can possibly result in two alternative evolutionary fates: i) neofunctionalization where one copy acquires an entirely new function whereas the other copy maintains the original function. ii) Subfunctionalization, in which each copy adopts part of the task of their parental gene (Ohno,1970; Nowak et al., 1997; Jenesen,1976; Orgel,1977; Hughes,1994). However, subfunctionalization is reported as a more prevalent outcome than neofunctionalization in nature. In any case, functional divergence of such paralogous proteins is found to be the key force shaping molecular network in organisms (Ohno, 1970). Recent studies also suggest that duplicate genes diverge mostly through the partitioning of gene expression as in subfunctionalization (Force et al.,1999; Wagner,2000; Gu et al.,2002). In addition, subfunctionalization can also take place at the protein function level leading to functional specialization, when one of the duplicated genes becomes better at performing one of the original functions of the progenitor gene (Hughes, 1994; Gu et al.,2002; Conant & Wolfe, 2008; Hughes, 1999; Zhang et al., 2002).

Myo-inositol-1-phosphate synthase (MIPS;EC5.5.1.4) is an evolutionary conserved enzyme which catalyzes the rate limiting step in well conserved inositol biosynthetic pathway and is extremely widespread in living organisms including plants (Loewus & Murthy, 2000; Majumder et al., 2003). The evolution of MIPS gene/ protein among the prokaryotes seems to be more divergent and complex than amongst the eukaryotes, however they preserve a conserve core catalytic domain among the MIPS proteins (Majumder et al., 2003).

Many of the plant species are known to contain more than one copy of gene encoding MIPS and are hypothesized to arise through gene duplication. Expression studies of multiple gene encoding MIPS have revealed the possibility of specialized role for individual enzyme isoforms. Previously, two genes encoding MIPS have been identified and characterized from chickpea by Kaur *et al.* A comparative study of two divergent genes (*CaMIPS1* & *CaMIPS2*)

reveals features of both functional redundancy and diversification (Kaur et al., 2008). This chapter explores how a possible gene duplication of MIPS gene in chickpea lead to a functional diversification that perhaps contributed adaptive evolution to the plant.

## 2. Gene duplication and functional divergence of MIPS in chickpea

### 2.1 Evolution and diversification of MIPS

The inositols are the nine isomeric forms of cyclohexane hexitols and *myo*-inositol is the most abundant and physiologically favored molecule in the biological system.

The biosynthesis of *myo*-inositol has been acknowledged as an evolutionary conserved pathway and its importance across biological organisms from different domains of life has been recognized for long time. The first and rate limiting step of this pathway is catalyzed by an evolutionary conserved enzyme named as MIPS.

MIPS particularly catalyzes the conversion of glucose 6- phosphate (Glc6P) to *myo* inositol 1- phosphate (Ins1P) through an internal oxidoreduction reaction involving NAD<sup>+</sup>. Subsequently inositol 1- phosphate is dephosphorylated to produce free inositol by Mg<sup>2+</sup> dependent *myo*- inositol 1- phosphate phosphatase (IMP: EC 3.1.3.25) (Loewus & Loewus,1983;Loewus,1990). This free *myo* inositol occupies the central position in inositol metabolism since this free inositol can be channellized to various metabolic routes and produce different inositol derivatives (Fig-1) (Loewus & Murthy, 2000; Loewus,1990).

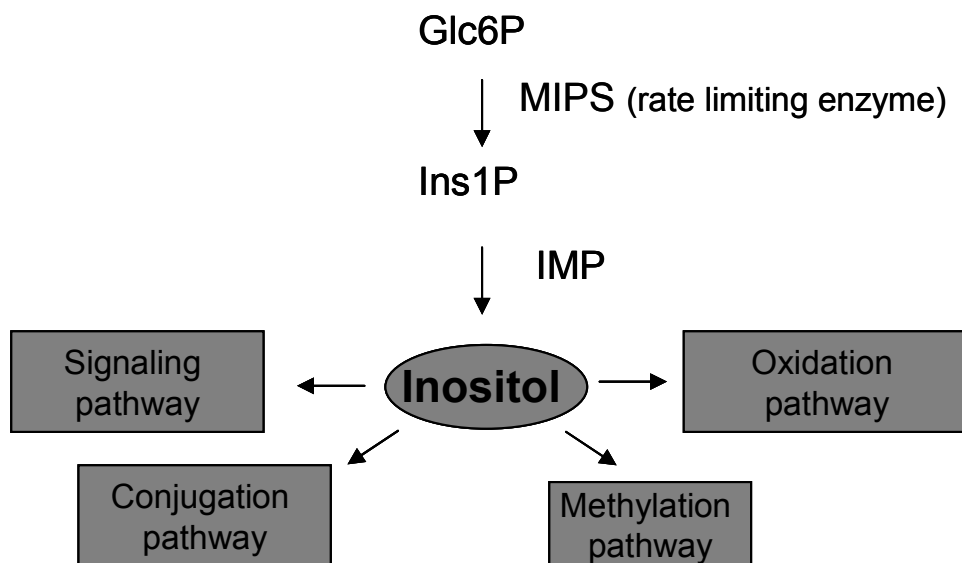


Fig. 1. Inositol biosynthesis and its consumption in other pathway.

This free inositol and its derivatives have acquired diverse functions over the course of evolution. As for example, inositol containing phospholipids are the important constituents of many archaea. Few thermophilic archaea also use inositol phosphodiester as thermo protective solutes. Then with the emergence and diversification of eukaryotes, function of

inositol and its derivatives proliferated dramatically. So far, inositol and its derivatives have been shown to be involved in growth regulation, membrane biogenesis, hormone regulation, signal transduction, pathogen resistance and stress adaptation in higher plants (Loewus & Murthy, 2000; Stevenson et al., 2000; Michell, 2008).

Since the usage and distribution of inositol and inositol derivatives are reported in all domains of life, it is imperative to contain MIPS enzyme in diverse organisms such as archaea, eubacteria, parasites, animals, higher plants and many others. Few higher plants and algae are reported to have both cytosolic and chloroplastic isoforms of MIPS. However, the biochemical and enzymatic properties of these two forms do not differ significantly between each other. Recent studies suggest that rice chloroplastic MIPS is coded by *OsINO1-1* gene located on chromosome 3 (RayChaudhuri et al. 1997; Ray et al., 2010).

The structural gene coding (*INO1*) for this ancient enzyme was first identified and cloned in *Saccharomyces cerevisiae* (Donahue & Henry, 1981). Subsequently more than 80 *INO1* genes were reported from various sources including both prokaryotes and eukaryotes.

Evolution and diversification of MIPS has been highlighted by Majumder et al. (2003) and a clear difference between prokaryotic and eukaryotic MIPS protein sequences was observed when compared among each other. The MIPS protein sequences of prokaryotes are quite divergent among themselves and significantly distinct than any other known eukaryotic sequences. In contrast, the eukaryotic MIPS sequences show remarkable similarities among each other. A phylogenetic tree constructed to include few representative MIPS sequences from diverse organisms present an overall evolutionary divergence of this enzyme in the biological kingdom. The higher plants constitute one close subgroup, while the higher animals, protozoa, fungi form the other subgroups in the eukaryotic cluster (Fig-2). In *Archeoglobus*, MIPS shows more sequences similarity to the eukaryotic MIPS than the other known prokaryotic ones and thereby all eukaryotic MIPS seems to have evolved from one common stock, probably from the fusion of an archaebacterial and eubacterial MIPS genes (Fig-2 & 3).

Four stretches of amino acid residues (GWGGNG, LWTANTERY, NGSPQNTFVPGL and SYNHLGNNDG) are found to be conserved in MIPS proteins of all eukaryotes and among them; SYNHLGNNDG is identified as highly conserved. Interestingly among higher plants, MIPS enzyme shows greater conservation in addition to these four domains (Fig-3). Many of the plant species also possess multiple genes encoding MIPS and are thought to arise through gene duplication in course of time.

Subsequent analysis of crystal structures of various MIPS proteins provide ample evidence towards the presence of conserved "core structure" in all MIPS proteins throughout evolution. Moreover, some of the important amino acid residues are identified in the active site of the yeast MIPS and are shown to be highly conserved in all eukaryotic MIPS. These amino acids are considered to be the part of a "eukaryotic core structure" which has remained largely the same during evolution, despite the divergence in rest of the sequences over time (Fig-3) (Stein & Geiger, 2002; Norman et al., 2002).

Crystal structure analysis of MIPS from *Saccharomyces cerevisiae* also revealed that each monomer of the homo-tetrameric MIPS has three functionally important structural domains namely the NAD binding Rossman fold, the catalytic binding site and the core domain. This study also exemplifies a case of induced fit model for binding of the substrate with the catalytic domain of the enzyme. (Stein & Geiger, 2002)

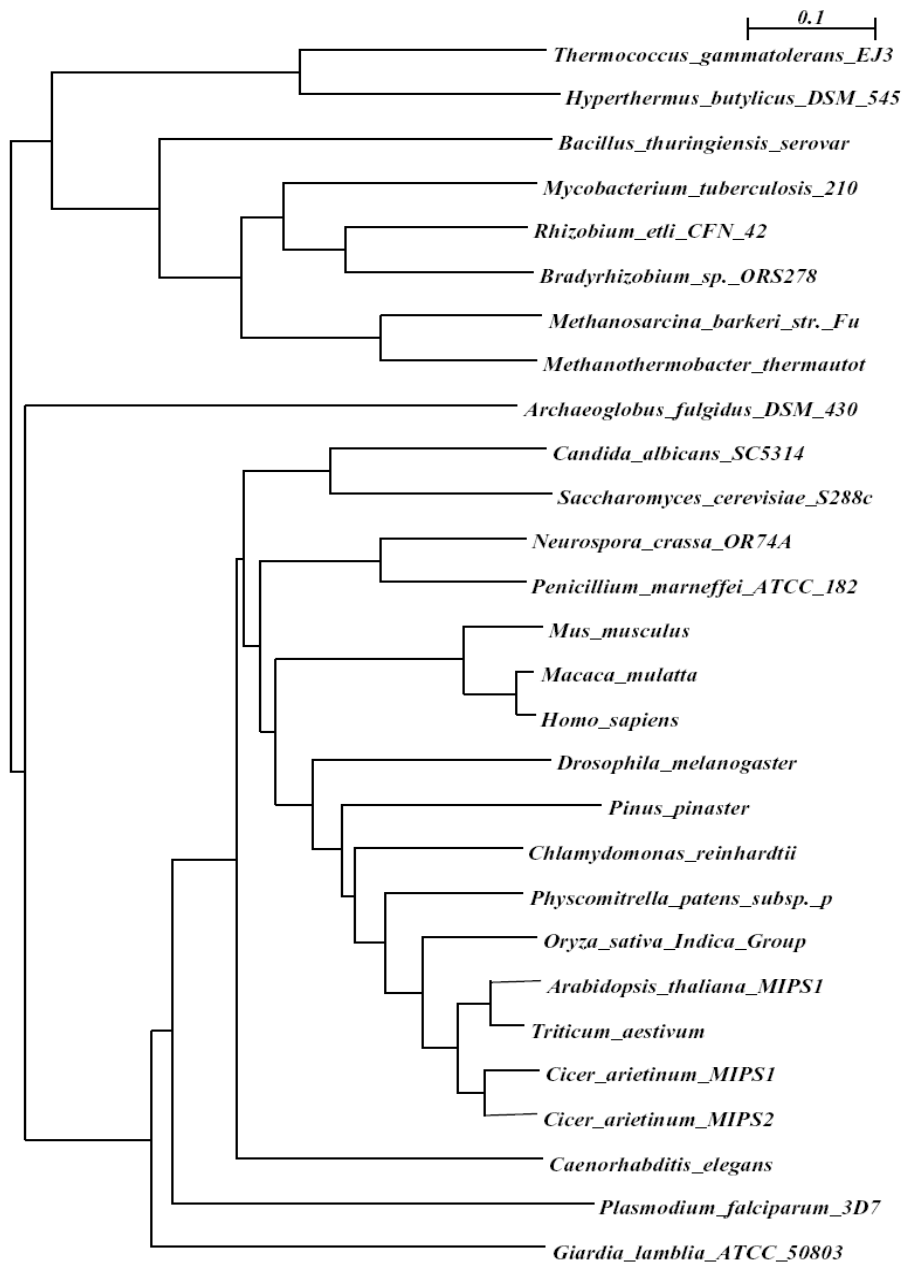


Fig. 2. A phylogenetic tree of few representative MIPS amino acid sequences from various domains of living organisms. Neighbour-Joining algorithm was used to construct tree from the distance matrix using Clustal X. Thousand rounds of bootstrapping were performed to ensure the validity of the tree.

```

CaMIPS1      -----MFIENFKVDSPNVKYTETEIQSVVNYETTEL VHENRNGTYQWIVVKPKTVKYEFK
CaMIPS2      -----MFIESFKVESPNVKYTDTEIQSVVSYETTEL VHENRNNTYQWIVVKPKTIKYEFK
OsMIPS       -----MFIESFRVESPHVRYGAAEIESDYQYDTELVHESH DGASRWIVRPKSVRYNFR
HsMIPS       -----MEAAAQFFVESPDVVYGPEAIEAQYEYRTRVSREG-----GVLKVHPTSTRFTFR
ScMIPS       MTEDNIAPITSVKVVTDKCTYKDNELLTKYSYENAVVTKTAS---GRFDVTPTVQDYVFK
PfMIPS       -----
MtMIPS       -----MSEHQ S
AfMIPS       -----

```

```

CaMIPS1      TDTHVP-KLGVMLV GWGGNNGSTLTGGVIANREGISWATKDN IQQANYFGSLTQASATRV
CaMIPS2      TQTHVP-KLGVMLV GWGGNNGSTLTGGVIANREGISWATKDN IQQSNYFGSLTQASATRV
OsMIPS       TTTTVP-KLGVMLV GWGGNNGSTLTAGVIANREGISWATKDKVQQANYFGSLTQASTIRV
HsMIPS       TARQVP-RLGVMLV GWGGNNGSTLTAAVLANRLRLSWPTRSGRKEANYFGSLTQAGTVSL
ScMIPS       LDLKKPEKLGIMLI GLGGNNGSTLVASVLANKHNVEFQTKGEVKQPNYFGSMTQCSTLKL
PfMIPS       -----MVRVAIIQGQYVASIFAVGLERIKE-----GELGYG
MtMIPS       LPAPEASTEVRVAIVGVGNCASSLVQGV EYYN-----ADDTSTVPG-----
AfMIPS       -----MKVWLVGAYGIVSTTAMVGARAIERGIAPKIGLVSELPHFEG-----

```

: . \*

```

CaMIPS1      GSFQ-GEEIYAPFKSLLPMVNP DDIVGGWDI SDMNLADAMARA-RVFDIDLQKQLRPYM
CaMIPS2      GSFQ-GEEIYAPFKSLLPMVNP DDIVGGWDI NNMNLADAMGRA-RVFDIDLQKQLRPYM
OsMIPS       GSYN-GEEIYAPFKSLLPMVNPDDL VFGGWDISNMNLADAMTRA-KVL DIDLQKQLRPYM
HsMIPS       GLDAEGQEVFVFPFSAVLPMVAPNDLVFDGWDISSNLAEAMRRA-KVL DWDGQEQLWPHM
ScMIPS       GIDAEGNDVYAPFNSLLPMVSPNDFVVGWDINNADLYEAMQRS-QVLEYDLQRLKAKM
PfMIPS       -----IPLANELPIKVEDIKIVASYDV DKTIGLPLSEI-VQRYWKGNVPESLQE
MtMIPS       -----LMHVRFGPYHVRDVKFVAADFVD AKKVGFDLSDA-IFASENNTIKIADVA
AfMIPS       -----IEKYAPFSFEFGGHEIRLLSNAYEAAKEHWELNRHFDREILEAVKSDL

```

\* . .

```

CaMIPS1      ESMVPLPGIYDPDFIAANQGDRANNVI KGTKR-----EQ INQIIKDIKEFK EAKV
CaMIPS2      ESMVPLPGIYDPDFIAANQGDRANNVI GTKK-----EQ LQQIIKDIKEFK EASKI
OsMIPS       ESMVPLPGIYDPDVIAANQGSRRANNVI GTKK-----EQ MEQIIKDIREFKEKSKV
HsMIPS       EALRPRPSVYIPEFIAANQSARADNLI PG SRA-----Q QLEQIRRDIRDFRSSAGL
ScMIPS       SLVKPLPSIYYPDFIAANQDERANNCINLDEKGNVTTRGKWHLQRI RRD IQNFKEENAL
PfMIPS       VFVRKGIIHLGSLRNLPIEATGLEDEMT-----L KEAIERLVEEWKEKKVD
MtMIPS       PTNVIVQRGPTLDGIGK-----Y YADTIELSDAEPV
AfMIPS       EGIVARKGTALNCGSGIKELGDIKTLEGEGLS-----L AEMVSRIEEDIKSFAD

```

: .

```

CaMIPS1      DRVVVLWTANTERY SNLVVGLNDTMENLFAAVDRNE-SEISPSTLFAIACV TENVPFING
CaMIPS2      DKVVVLWTANTERY SNVVGLNDTMENLLASVDKNE-AEISPSTLYALACV TENVPFING
OsMIPS       DKVVVLWTANTERY SNVCVGLNDTMENLLASVDKNE-AEISPSTLYAIACVMEGIPFING
HsMIPS       DKVIVLWTANTERFCEVIPGLNDTAENLLRTIELG--LEVSPSTLFAVASILEGCAFLNG
ScMIPS       DKVIVLWTANTERY VEVSPGVNDTMENLLQSIKNDH-EEIAPSTIFAASILEGVPIYNG
PfMIPS       VIINVPTTEAFTPFKGLEELEKAIKDNNKERLTATQ-AYAYAAAQYAKE--VGGA AFVNA
MtMIPS       DVVQALKEAKVDVLVSYPVGSEE-----ADKFYAQCAIDAGVAFVNA
AfMIPS       DETVVINVASTEPLPNYSEEHSGSLEGFERMIDEDRKEYASASMLY AYAALKLGLPYANF

```

. . .: \* . .: \*

```

CaMIPS1      SP-QNTFVPGLIDLAIKRNTLIGGDDFK--SGQTKMKSVLVDFLVGAGIKPTSIVSYNHL
CaMIPS2      SP-QNTFVPGLIDLAIQRNSLIGGDDFK--SGQTKMKSVLVDFLVGAGIKPTSIVSYNHL
OsMIPS       SP-QNTFVPGLIDLAIKNNCLIGGDDFK--SGQTKMKSVLVDFLVGAGIKPTSIVSYNHL
HsMIPS       SP-QNTLVPGALELAWQHRVFVGDDDFK--SGQTKVKSVLVDFLIGSGLKTMSIVSYNHL
ScMIPS       SP-QNTFVPGLVQLAEHEGTFIAGDDLK--SGQTKLKSVLQAQLVDAGIKPVSIASYNHL
PfMIPS       IPTLIANDPAFVELAKESNLVIFGDDGA--TGATPLTADILGH LAQRNRHVLDIVQFNIG
MtMIPS       LPVFIASDPVWAKKFTDAGVPIVGDDIKSQVGATITHRV LAKLFEDRGVQLDRTMQLNVG

```

AfMIPS	TPSPGSAIPALKEAEKKGVPHAGNDGK--TGETLVKTTLAPMFAYRNMEVVGWMSYNIL * : * . *: * : : . *
CaMIPS1	GNNDGMNLSAPQTFRSKEISKSNVDDMVNSNGILY--APGEHPDHVVVIKYVPYVGDSE
CaMIPS2	GNNDGMNLSAPQTFRSKEISKSNVDDMVNSNAILY--QPGEHPDHVVVIKYVPYVADSK
OsMIPS	GNNDGMNLSAPQTFRSKEISKSNVDDMVSSNAILY--ELGEHPDHVVVIKYVPYVGDSK
HsMIPS	GNNDGENLSAPLQFRSKEVSKSNVDDMVQSNPVLY--TPGEEPDCVVIKYVPYVGDSK
ScMIPS	GNNDGYNLSAPKQFRSKEISKSSVIDDIASNDILYNDKLGKKVDHCIVIKYMKPVGDSK
PfMIPS	GNTDFLALTDKERKNSKEYTKSSVVEDILG-----YDAPHFIKPTGYLEPLGDKK
MtMIPS	GNMDFLNLMLERERLESKKISKTKAVTSNLKR-----EFKTKDVHIGPSDHSVGLDDRK
AfMIPS	GDYDGGKVLSDARNKESKVLSDKVKLEKMLG-----YSPYSITEIQYFPSLVDNK *: * : . ** : * . : : : : * *
CaMIPS1	RAMDEYTSSEIFMGGKSTIVLHNTCEDSLLAAPIILDVLVLAELSTRIQFKSEAE-----EN
CaMIPS2	RAMDEYTSSEIFMGGKSTIVLHNTCEDSLLAAPIILDVLVLAELSTRIQFKSQH-----ED
OsMIPS	RAMDEYTSSEIFMGGKSTIVLHNTCEDSLLAAPIILDVLVLAELSTRIQLKAE-----EE
HsMIPS	RALDEYTSSEIMLGSTNTIVLHNTCEDSLLAAPIMLDLALLTELQCRVSFCTDM-----DP
ScMIPS	VAMDEYYSSEIMLGHNRIISIHNVCEDSLLATPLIIDLLVMTEFCTRVSYKKVDPVKEDAG
PfMIPS	FIAMHIEYISFNGARDELIIAGRINDSPALAGLLVDLARLGKIAVDK-----K
MtMIPS	WAYVRLEGRAFGDVPLNLEYKLEVWDSNSAGVIIDAVRAAKIAKDRGIG-----
AfMIPS	TAFDFVHFKGFLGKLMKFYFIWDAIDAIVAAPLILDIARFLFAKKKGVKG : . : *: : : * : .
CaMIPS1	KFHTFHPVATILSYLTKAPLVPPGTPVNVNALSQRAMLENIMRACVGLAPENNMILEYK-
CaMIPS2	KFHSFHPVATILSYLTKAPLVPPGTPVNVNALSQRAMLENIMRACVGLAPENNMILEYK-
OsMIPS	KFHSFHPVATILSYLTKAPLVPPGTPVNVNALSQRAMLENIMRACVGLAPENNMILEYK-
HsMIPS	EPQTFHPVLSLLSFLFKAPLVPPGSPVNVNALSQRAMLENIMRACVGLPQQNHMLLEHKM
ScMIPS	KFENFYPLVTLFLSYWLKAPLTRPGFHPVNLNQRALLENFLRLILGLPSQNELRFEERL
PfMIPS	---EFGTVYPVNAFYMKNPGPKEAKNIPRIIAYEKLQWAGLPPRYL-----
MtMIPS	-----GPVIPASAYLMKSPPEQLPDDIARAQLEEFIIIG-----
AfMIPS	-----VVKEMAFFFSMPMDTNVINTHEQFVVLKEWYSNLK-----
CaMIPS1	-----
CaMIPS2	-----
OsMIPS	-----
HsMIPS	ERPGPSLKRVGPAATYPMNLKKGPVPAATNGCTGDANGHLQEPPMPTT
ScMIPS	L-----
PfMIPS	-----
MtMIPS	-----
AfMIPS	-----

Fig. 3. Multiple sequence alignment of MIPS from prokaryotes and eukaryotes. Proposed common active site amino acid residues for the *Mycobacterium tuberculosis* and *Saccharomyces cerevisiae* MIPS sequence are highlighted in green color and four conserved domains of eukaryotes (GWGGNG, LWTANTERY, NGSPQNTFVPGL and SYNHLGNNDG) are highlighted in yellow color. 43 variant positions between CaMIPS1 and 2 have been highlighted in blue color.

## 2.2 MIPS from chickpea: A case of functional divergence

Chickpea is an annual self-pollinated diploid legume crop which is mostly grown in the arid and semi arid regions of the world. Long term evolution and adaptation to harsh conditions make chickpea rich in resistant genes for environmental stresses including drought and cold. Several classes of genes controlling potential resistance have been

identified through genomic and proteomic studies (Ahmaed et al., 2005; Mantri ,2007; Bhusan et al., 2007).

In this particular plant, inositol seems to play an important role in drought tolerance besides growth and development, since inositol content and MIPS transcript was found to be significantly increased under dehydration condition (Boominathan et al.,2004). Subsequently, chickpea is reported to have two MIPS coding genes (*CaMIPS1* and *CaMIPS2*) (Kaur et al. 2008) and both genes are revealed to have overall similar structure consisting of 9 introns and 10 exons (Fig-4). Sequence analysis of these two genes show high similarity (>85%) in their coding regions but their non-coding or 5' and 3' flanking regions are extremely divergent. Moreover length of each exon is similar between these two genes while the size of introns varies. Such findings suggest that these two MIPS genes most likely arose by ancestral gene duplication and have undergone considerable sequence divergence.

In spite of the remarkable resemblance in their coding sequences, some base substitutions occurs in exons leading to changes in 43 amino acids in protein sequences, however, maintaining four highly conserved functional domains and known active site amino acids of MIPS (Fig-3) (Majumder *et al.* 2003). Among these 43 amino acids, 19 amino acids differ considerably between *CaMIPS1* and *CaMIPS2* while rest of the amino acid substitutions are relatively insignificant, i.e. substitution between amino acids having similar physico chemical properties (Kaur et al., 2008).

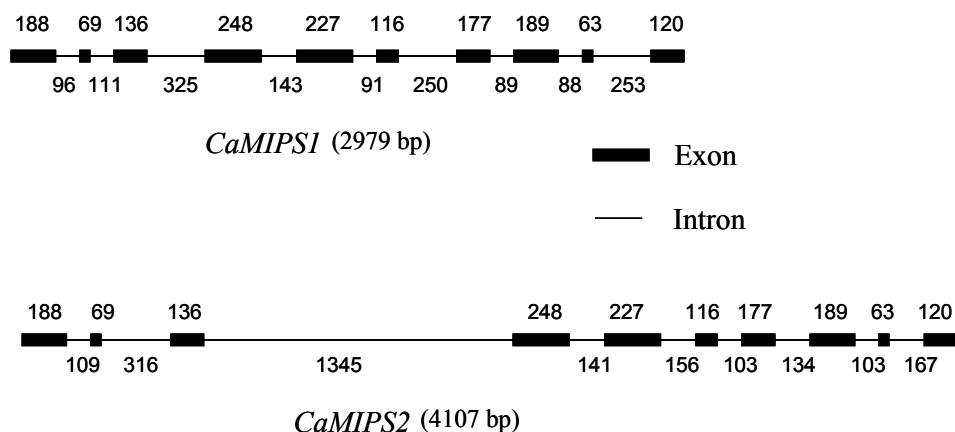


Fig. 4. Diagrammatic representation of *CaMIPS1* and *CaMIPS2* genomic structure. Length of exon and intron indicated in bp. [Modified from Kaur et al., 2008]

Functional divergence after gene duplication can result in following alternative fates: One copy acquires a novel function (neofunctionalization) or one copy loses its function completely or each copy adopts part of the task of their parental gene (subfunctionalization) (Ohno, 1970; Nowak et al., 1997; Jenesen, 1976; Orgel,1977;Hughes,1994).

Functional complementation and in-vitro enzymatic properties were analyzed to check the fate of these two genes. First to check the functional identity of these two divergent genes, a complementation experiment was carried out in natural inositol auxotroph *Schizosaccharomyces pombe* PR109 which clearly demonstrates that both *CaMIPS1* and

*CaMIPS2* indeed encode functional MIPS enzymes. Subsequently, the enzymatic properties of these two enzymes were examined since *CaMIPS1* and *CaMIPS2* polypeptides are reported to have some differences in their amino acid sequences.

Both enzymes showed nearly same  $K_m$  values for Glc6P suggesting the similar substrate specificity. For both proteins, the optimum temperature for enzyme activity is at 35°C and optimum pH is 7.0 suggesting the similar biochemical characteristics (table1).

Further the enzymatic activities of each protein under stress environment in *invitro* conditions were examined and the activities of these two enzyme proteins were shown to differ significantly in response to high temperature and salt concentration (Kaur et al., 2008). *CaMIPS1* activity is considerably affected at high temperature or in presence of increasing sodium chloride concentration while the *CaMIPS2* activity is less affected in similar conditions and thereby retaining higher activity than *CaMIPS1* (Fig-5).

The amino acid substitutions in protein sequence as analyzed by sequence comparison and the higher enzyme activity in *CaMIPS2* under stress condition also indicates that it might be evolved during the course of time to function better under stress conditions. This differential activity towards high temperature and salt of these two enzymes could be supported by the bioinformatics analysis in respect to the available yeast MIPS crystal structure and salt tolerant *PcINO1* (MIPS coding gene from Salt tolerant *Porteresia coarctata*) protein sequence (Majee et al., 2004). Based on the bioinformatics study, *CaMIPS2* appears to be more stable towards destabilizing factors such as high temperature, salt, etc, thereby retains better functionality under such conditions (Kaur et al., 2008). Subsequently, growth pattern of *CaMIPS1* & *CaMIPS2* transformed *Schizosaccharomyces pombe* under stress conditions were analyzed and *CaMIPS2* transformed *S. pombe* cells were reported to grow or survive better than *CaMIPS1* transformants both at high temperature and salt environment (Fig-6) suggesting *CaMIPS2* gene product functions more efficiently under stress conditions due to its stress tolerant property and hence provide sufficient inositol to grow as compared to *CaMIPS1*.

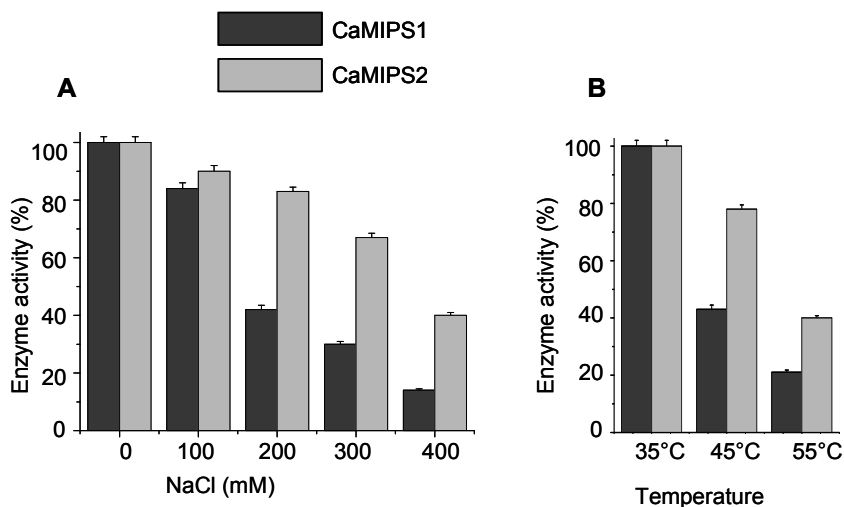


Fig. 5. Effect of salt (A) & temperature (B) on *CaMIPS1* and *CaMIPS2* enzyme activity. [Modified from Kaur et al., 2008]



Characters	CaMIPS1	CaMIPS2
Km		
Gluc 6-P	2.63 mM	2.70 mM
NAD <sup>+</sup>	0.181 mM	0.192 mM
Vmax		
Gluc 6-P	0.074 μmole min <sup>-1</sup>	0.075 μmole min <sup>-1</sup>
NAD <sup>+</sup>	0.069 μmole min <sup>-1</sup>	0.070 μmole min <sup>-1</sup>
pH optima	7.5	7.5
Temp. optima	35°C	35°C

[Modified from Kaur et al., 2008]

Table 1. Biochemical characterization of recombinant CaMIPS1 and CaMIPS2 enzymes.

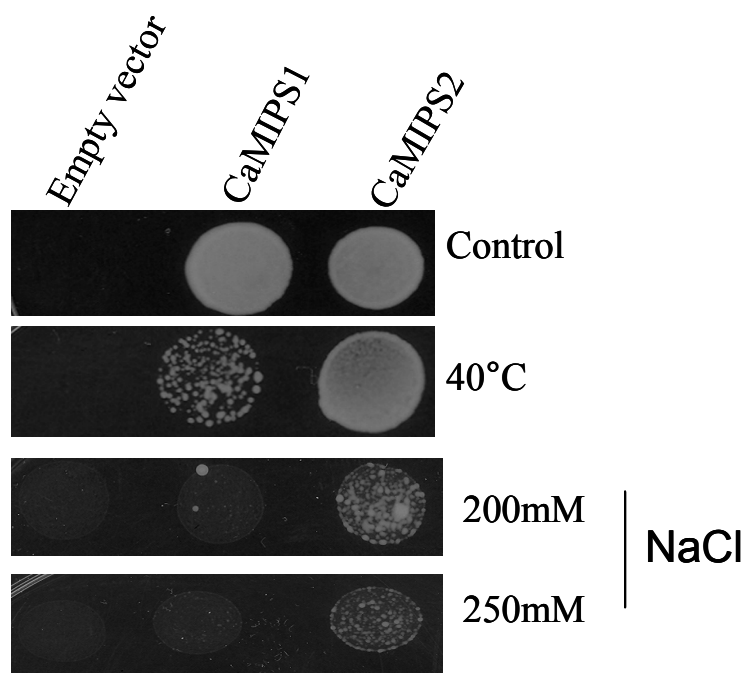


Fig. 6. Growth pattern of *Schizosaccharomyces pombe* transformed with CaMIPS1 and CaMIPS 2 at high temperature and salt environment. [Modified from Kaur et al., 2008]

Recent studies suggest that duplicate genes diverge mostly through the partitioning of gene expression as in subfunctionalization and thereby being expressed in a differential manner; redundant genes may acquire functional divergence (Force et al., 1999; Wagner, 2000; Gu et al., 2002). This hypothesis was examined on *CaMIPS1* and 2. *CaMIPS1* gene was shown to express in root, shoot, leaves, and flower in fairly equal abundance but no transcript was observed in seed, while *CaMIPS2* transcript was observed

in all examined tissues including seed. This result proposes that *CaMIPS1* and *CaMIPS2* genes are indeed differentially regulated in different organs to coordinate inositol metabolism with cellular growth as hypothesized previously (Loweus & Murthy, 2000). Subsequently, expression pattern of these two genes are examined in various environmental stresses. Interestingly, *CaMIPS2* was shown to be induced at different level in various environmental stresses while level of *CaMIPS1* transcript was found to be unaltered by such stresses (Fig-7). This differential expression is also supported by the divergence of their upstream regulatory sequences.

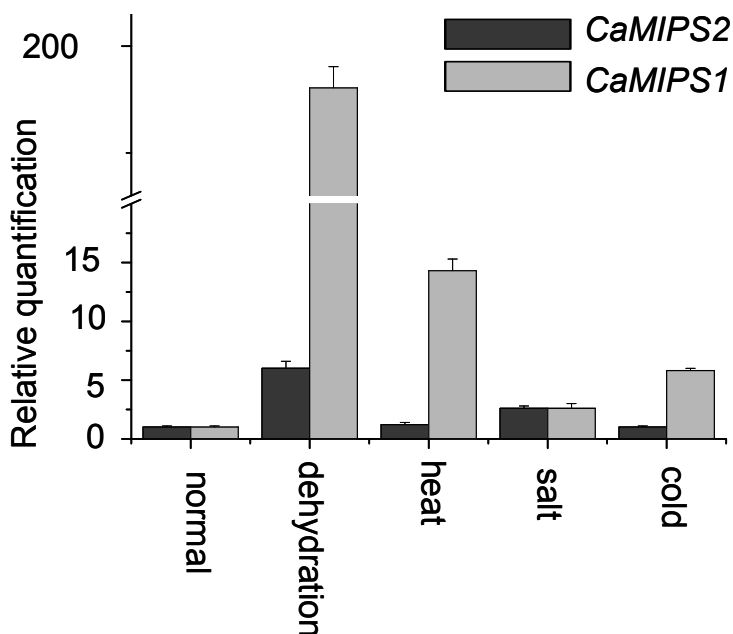


Fig. 7. Expression analysis of *CaMIPS1* and *CaMIPS2* through real time PCR analysis under various stresses. [Modified from Kaur et al., 2008]

### 3. Conclusion

Gene duplication, followed by sequence divergence leads to functional divergence of the paralogous proteins, is a major force for adaptation of living organisms. Without gene duplication, the plasticity of genome or organism in adapting to changing environment would be very limited. Chickpea plants are known to be evolved and diversified considerably over time and acquired subsequently various potential genes for their adaptation to environmental stresses. It seems that this drought tolerant legume plant requires more inositol for their adaptation particularly under drought condition and hence acquired *CaMIPS2* over time. Collectively, our results exemplified that *CaMIPS1* and *CaMIPS2* are differentially expressed in chickpea to play discrete though overlapping roles in plant;

however *CaMIPS2* is likely to be evolved through gene duplication, followed by adaptive changes in its sequences to function better under environmental stresses and thereby play a key role in environmental stress adaptation along with other aspects of inositol metabolism in chickpea.

#### 4. Acknowledgement

This work was supported by a grant from Department of Biotechnology (Next Generation Challenge Programme on Chickpea Genomics), Department of Science and Technology (Fast Track Scheme), Government of India.

We also like to acknowledge the Research support from National Institute of Plant Genome Research, New Delhi. H.K. thanks the Council of Scientific and Industrial Research, Government of India, for Senior Research Fellowship.

#### 5. References

- Ahmad, F.; Gaur, P. M. & Croser, J. (2005) Chickpea (*Cicer arietinum* L.) In genetic resources, Chromosome Engineering and Crop Improvement. Grain legumes Edited by Singh R, Jauhar P. vol1, pp 187-217, CRC Press. USA
- Bhusan, D.; Pandey, A.; Choudhary, M.K.; Datta, A.; Chakraborty, S. & Chakraborty, N. (2007) Comparative proteomics analysis of differentially expressed proteins in chickpea extracellular matrix during dehydration stress. *Molecular & Cellular Proteomics*, vol 6:1868-1884.
- Boominathan, P.; Shukla, R.; Kumar, A.; Manna, D.; Negi, D.; Verma, P.K. & Chattopadhyay, D. (2004) Long term transcript accumulation during the development of dehydration adaptation in *Cicer arietinum*. *Plant Physiology*, vol 135: 1608-1620
- Conant, G.C. & Wolfe, K.H. (2008) Turning a hobby into a job: How duplicated genes find new functions. *Nature Review Genetics*, vol 9: 938-950
- Donahue, T.F. & Henry, S.A. (1981) Myo-inositol 1- phosphate synthase. Characteristics of the enzyme and identification of its structural gene in yeast. *Journal of Biological Chemistry*, vol 256: 7077-7085
- Flagel, L.E. & Wendel, J.F. (2009) Gene duplication and evolutionary novelty in plants. *New Phytologist*, vol 183: 557-564.
- Force, A.; Lynch, M.; Pickett, F.B.; Amroes A.; Yan, Y.-L. & Postlethwait, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, vol 151, 1531-1545
- Gu, Z.; Nicolae, Lu, H.-S. & Li, H.W. (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in Genetics*, vol 18, 609-613
- Hughes, A.L. (1994) The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society, Lond Ser. B* 256: 119- 124
- Hughes, A.L. (1999) *Adaptive Evolution of genes and genomes*, Oxford University press.
- Jensen, R.A. (1976) Enzyme recruitment in the evolution of new function. *Annual Review of Microbiology*, vol 30, 409-425
- Kaur, H.; Shukla, R.; Yadav, G.; Chattopadhyay, D. & Majee, M. (2008) Two divergent genes encoding L -myo inositol 1 phosphate synthase1 (*CaMIPS1*) and 2 (*CaMIPS2*) are differentially expressed in chickpea. *Plant, Cell & Environment*, vol31, 1701-1716

- Loewus, F.A. & Loewus, M.W. (1983) *Myo*- inositol: its biosynthesis and metabolism. *Annual Review of Plant Physiology*, vol 34,137-161
- Loewus, F.A. & Murthy, P.N. (2000) *Myo*- inositol metabolism in plants. *Plant Science*, vol 150, 1-19
- Loewus, F.A. (1990) *Inositol biosynthesis*. In *Inositol metabolism in plants* (Morre D.J., Boss W.F. & Loewus F.A. eds). pp13-19, New York, Wiley-Liss., USA
- Majee, M.; Maitra, S.; Dastidar, K.G.; Pattnaik, S.; Chatterjee, A.; Hait, N.C.; Das, K.P. & Majumder, A.L. ( 2004) A novel salt tolerant L *myo*- inositol -1-phosphate synthase from *Porteresia coarctata* (Roxb) Tateoka, a halophytic wild rice. *Journal of Biological Chemistry*, vol 279, 28539-28552
- Majumder, A.L.; Chatterjee, A.; Dastidar, K.G. & Majee, M. (2003) Diversification and evolution of L- *myo*-inositol 1 phosphate synthase. *FEBS Letter* ,vol 533, 3-10.
- Mantri, N.L.; Ford, R.; Coram, T.E. & Pang, E.C. (2007) Transcriptional profiling of chickpea genes differentially regulated in response to high salinity, cold, and drought. *BMC genomics*, vol 8:303
- Michell, R.H. (2008) Inositol derivatives: evolution and functions. *Nature reviews Molecular Cell Biology* ,vol9, 151-161
- Norman, R.A.; McAlister, M.S.B.; Murray Rust, J.; Movahedzadeh, F.; Stoker, N.G. & McDonald, N.Q. (2002) Crystal structure of inositol 1 phosphate synthase from *Mycobacterium tuberculosis*, a key enzyme in phosphatidyl inositol synthesis. *Structure* ,vol10: 393-402
- Nowak, M.A.; Boerlijst, M.C.; Cooke, J, Smith, M.J. (1997) Evolution by genetic redundancy. *Nature*, vol 388,167-171
- Ohno, S. (1970) *Evolution by gene duplication*. Springer -Verlag. New York,, USA
- Orgel, L.E. (1977) Gene duplication and the origin of proteins with novel functions. *Journal of Theoretical Biology*, vol 67,773
- Ray, S.; Patra, B.; Das Chatterjee, A.; Ganguli, A. & Majumder, A.L. (2010) Identification and organization of chloroplastic and cytosolic L-*myo*-inositol 1- phosphate synthase coding gene (s) in *Oryza sativa*: comparison with the wild halophytic rice, *Porteresia coarctata*. *Planta* 231:1211-1227
- RayChaudhuri, A.; Hait, N.C.; DasGupta, S.; Bhaduri, T.J.; Deb, R. & Majumder, A.L. (1997) L *myo*- inositol 1-phosphate synthase from plant sources (characteristics of the chloroplastic and cytosolic enzymes). *Plant Physiology* 115: 727-736
- Stein, A.J. & Geiger, J.H. (2002) The crystal structure and mechanism of L *myo* inositol 1 phosphate synthase. *Journal of Biological Chemistry* 277: 9484-9491.
- Stevenson, J.M.; Perera, I.Y.; Heilmann, I.; Persson, S. & Boss, W.F. (2000) Inositol signaling and Plant Growth. *Trends in Plant Science* 5: 252-258.
- Wagner, A. (2000) Decoupled evolution of coding region and mRNA expression patterns after gene duplication: Implication for the neutralist selectionist debate. *Proceedings of National Academy of Sciences USA*. vol 97: 6579-6584
- Zhang, J. (2003) Evolution by gene duplication; an update: *Trends in Ecology and evolution*, vol18, 292-298.
- Zhang, J.; Zhang, Y.P. & Rosenberg, H.F. (2002) Adaptive evolution of duplicated pancreatic ribonuclease gene in a leaf eating monkey. *Nature Genetics*, vol 30, 411-415

# On the Specialization History of the ADP-Dependent Sugar Kinase Family

Felipe Merino and Victoria Guixé  
*Laboratorio de Bioquímica y Biología Molecular*  
*Facultad de Ciencias*  
*Universidad de Chile*  
*Chile*

## 1. Introduction

Sugars are one of the most common carbon sources used by heterotrophic organisms. Indeed, sugar phosphorylation is thought to be a key step in the cellular metabolism since, just after transport into the cell, these molecules are phosphorylated to trap them for further metabolic processing. There are several known pathways used to produce pyruvate from the incoming sugar (like glucose or galactose) which is accompanied by the synthesis of ATP and the production of reductive power. Amongst them, the Embden-Meyerhof pathway, or glycolysis, seems to be the most commonly used. Some microorganisms can also use the Entner-Doudoroff pathway. Also, although the pentose phosphate pathway is generally associated with nucleotide synthesis and reductive power in the form of NADPH it also can be linked to the flux from glucose to pyruvate as this pathway has fructose-6-phosphate and glyceraldehyde-3-phosphate as intermediates. Some microorganisms, such as *Lactococcus lactis*, use a pathway very similar to glycolysis, but instead of start from glucose they use galactose as main carbon source. In this fashion glucokinases are replaced by galactokinases and phosphofructokinases by tagatose-6-phosphate kinases (van Rooijen et al., 1991).

Interestingly, all the above mentioned pathways ultimately converge through glyceraldehyde-3-phosphate. In this way, the main difference between them is what happens with the hexoses. Here, one of the most important reactions are the initial phosphorylations, e.g. phosphorylation of glucose, fructose-6-phosphate, galactose, tagatose-6-phosphate.

Early on the 90s it was already recognized that the transfer of the  $\gamma$ -phosphate of ATP to several sugars was catalyzed by at least three different non-homologous protein families: the hexokinase family, the ribokinase family, and the galactokinase family (Bork et al., 1993). The hexokinase family contains enzymes with wide specificities including glucokinases, ribulokinases, gluconokinases, xylulokinases, glycerokinases, fructokinases, rhamnokinases, and fucokinases (Bork et al., 1993). The galactokinase family contains enzymes that catalyze the phosphorylation of galactose, mevalonate, P-mevalonate, and homoserine (Bork et al., 1993). The ribokinase family on the other hand is very interesting since its members catalyze the transfer of the terminal phosphate of ATP to sugars like ribose, fructose, sugar containing molecules such as nucleosides, and sugar phosphate molecules like fructose-6-phosphate, fructose-1-phosphate, and tagatose-6-phosphate (Bork et al., 1993). This makes the ribokinase family the group with the broadest specificity amongst the above mentioned. It is clear that while the three groups share some similar substrates and hence are a great example of

convergent evolution the ribokinase family is the only one that contains enzymes able to phosphorylate sugar phosphates.

In particular, glucokinases have been extensively studied since they are in the top on many metabolic pathways, and hence some sort of metabolic hub, and also they are responsible for most of the flux control in glycolysis (Torres et al., 1988). On the other hand, while in normal conditions the phosphofructokinase from rat liver shows almost no control over the glycolytic flux, in starving conditions it becomes almost as important as glucokinase (Torres et al., 1988) which suggests that they become key in gluconeogenic conditions. Moreover, phosphofructokinases are extensively studied because they are highly regulated enzymes. In this light, phosphofructokinases have also been recognized as one of the key enzymes of glycolysis.

From the ribokinase family, one of the most studied enzyme is the phosphofructokinase-2 from *Escherichia coli* which is often referred to as a member of the PfkB subfamily (Cabrera et al., 2010). It is possible to find a second phosphofructokinase, called phosphofructokinase-1<sup>1</sup>, in the genome of *E. coli* which belongs to another family called PfkA. In this family, the most extensively studied members are the phosphofructokinase-1 from *E. coli* and the phosphofructokinase from *Bacillus stearothermophilus* (Evans et al., 1981; Schirmer & Evans, 1990). Initially it was thought that both PfkB and PfkA groups had a common origin (Wu et al., 1991), but now we know that they are two non-homologous families. Interestingly, while not phylogenetically related, both phosphofructokinase-1 and phosphofructokinase-2 from *E. coli* show strong inhibition at high concentrations of their substrate MgATP (Atkinson & Walton, 1965; Kotlarz & Buc, 1981), which suggests that this is a key requirement of this metabolic step. This reinforces the idea that these enzymes are strongly related to the balance between glycolysis and gluconeogenesis.

Indeed, it has been already demonstrated that the substrate inhibition is needed for the avoidance of a futile cycle of phosphorylation/dephosphorylation of fructose-6-phosphate/fructose-1,6-bisP which will ultimately lead to a net hydrolysis of ATP (Torres et al., 1997). Interestingly, some microorganisms present phosphofructokinases (also members of the PfkA family) which use polyphosphates as a source of phosphate and hence they do not appear to be regulated (Peng & Mansour, 1992).

## 2. Glucose degradation in the members of the Archaea domain

Nowadays, organisms can be classified in three principal domains of life: Bacteria, Eukarya, and Archaea (Woese & Fox, 1977; Woese et al., 1990). Interestingly, although there are some known archaea that growth in mesophilic conditions, most of them are extremophiles. Two main phylogenetic groups can be found inside Archaea: *Euryarchaeota* and *Crenarchaeota* (Allers & Mevarech, 2005). Also, recently based in environmental samples, two more groups called *Korarchaeota* and *Nanoarchaeota* has been proposed (Allers & Mevarech, 2005).

Considering the potential for technological applications, most of the attention has been directed to study those archaea able to grow in extreme temperature conditions (known as thermophiles or hyperthermophiles), extremely high salinities (known as halophiles), extremely low pH (known as acidophiles), and most commonly a combination of them. From the *Crenarchaeota*, the *Sulfolobus* and *Aeropyrum* genera receive lots of attention since both are aerobic thermophilic organisms. In the *Euryarchaeota*, the methanogenic organisms are

<sup>1</sup> Sometimes phosphofructokinase-1 and phosphofructokinase-2 from *E. coli* are called the major and minor enzyme respectively.

intensively studied. One of the most studied organism here is *Methanocaldococcus jannaschii*<sup>2</sup> (Jones et al., 1983) since it is one of the few organisms known to produce methane at extreme temperatures. Besides it, the *Halobacterium* and *Haloferax* genera are used as models for halophilic organisms while organisms from the *Thermococcus* and *Pyrococcus* genera are used as models of hyperthermophilic organisms. Here, by far, the most studied organism is *Pyrococcus furiosus*.

In these organisms, sugar degradation proceeds either through the Entner-Doudoroff or the Embden-Meyerhof pathway (Verhees et al., 2003). For instance, members of the *Thermoproteus*, *Thermoplasma*, and *Sulfolobus* genera degrade glucose through a modified version of the Entner-Doudoroff pathway where sugars are phosphorylated only at the 2-keto-3-deoxygluconate or glycerate level. While the former version is still able to produce one ATP molecule per glucose the later does not produce any ATP (for a review see Verhees et al. (2003)). On the other hand, up until the early 90s it was thought that some archaea of the *Euryarchaeota* used a modified unphosphorylated version of the Entner-Doudoroff pathway to degrade glucose (Mukund & Adams, 1991) which was called pyroglycolysis. However, in 1994 it was possible to demonstrate that, in fact, the flux to pyruvate proceeds through a highly modified version of the Embden-Meyerhof pathway (Kengen et al., 1994). Here, although all the intermediates are present, only four of the ten textbook enzymes are conserved (Verhees et al., 2003). In this pathway, the redox reactions are carried out by ferredoxin containing enzymes which latter use the electrons to reduce protons (producing hydrogen) to couple the proton motive force to ATP synthesis by means of a membrane bound hydrogenase enzyme (Sapra et al., 2003). Between the oxido-reductases present in these organisms, perhaps the most interesting is the glyceraldehyde-3-phosphate oxido-reductase. This enzyme is responsible for the single-step conversion of glyceraldehyde-3-phosphate to 3-phosphoglycerate in a phosphate independent manner (Mukund & Adams, 1995). Besides redox reactions, one of the most striking modifications seen in this version of the Embden-Meyerhof pathway is that the phosphorylation of glucose and fructose-6-phosphate is carried out by enzymes that use ADP and not ATP or polyphosphates as the phosphoryl donor (Kengen et al., 1994). These ADP-dependent enzymes are, in fact, homologous to each other and they show no sequence identity over the noise level with any of the hitherto known ATP, or polyphosphate dependent kinases (Tuininga et al., 1999). For this reason it was initially proposed that they belong to a new protein family called PfkC.

Given that these ADP-dependent enzymes were initially discovered in the hyperthermophilic archaeon *P. furiosus* (Kengen et al., 1994), it has been argued in the literature that the main reason for this "ADP-dependence" is the fact that ADP has a higher thermostability than ATP and also that both nucleotides are essentially equivalent since both have a similar standard  $\Delta G$  of hydrolysis. However, these arguments are highly misleading since, (i) as metabolism is a non-equilibrium process the free energy change upon phosphoryl transfers depends on the concentration of the metabolites, (ii) several ATP-dependent enzymes can be found in hyperthermophilic organisms, (iii) the ADP-dependent enzymes are also present in mesophilic organisms (see below), and (iv) the half life of ATP at high temperatures is higher than some other metabolic intermediates present in the Embden-Meyerhof pathway (Dörr et al., 2003).

The adaptive value of the appearance of the ADP-dependent enzymes has been a matter of great debate. As we have argued before (Guixé & Merino, 2009), it is most likely unrelated

<sup>2</sup> This organism was initially named *Methanococcus jannaschii* and was later renamed as *Methanocaldococcus jannaschii* to acknowledge the fact that those organisms from the *Methanococcus* genus are not thermophilic.

to the temperature at which most of the *thermococcales* grow. The most intriguing question arising here is what happens with the adenylate charge inside these archaea. As they present a glyceraldehyde-3-phosphate ferredoxin oxidoreductase (Mukund & Adams, 1995) which produces 3-phosphoglycerate in a single step that does not produce ATP and also considering both glucose and fructose-6-phosphate are phosphorylated using ADP as phosphoryl donor, it was thought that this modified glycolysis had a net ATP production of zero. However, it has been demonstrated by Sakuraba et al. (2004) that the pyruvate kinase from *P. furiosus* catalyze the synthesis of ATP from AMP, phosphoenolpyruvate, and Pi. In this way, the pathway from glucose to pyruvate produces two ATP molecules from every glucose molecule degraded.

Up until now, we have three protein families that contain phosphofructokinases: PfkA, the ribokinase family (which contains the PfkB-like kinases), and PfkC. While the first PfkA crystal structure (The phosphofructokinase from *B. stearothermophilus*) was solved in the 80s (Evans et al., 1981), the first PfkB-like crystal structure (the ribokinase from *E. coli*) (Sigrell et al., 1998) was solved in the late 90s, and the first PfkC crystal structure (The ADP-dependent glucokinase from *Thermococcus litoralis*) just in 2001 (Ito et al., 2001). As all of them were discovered before the middle 90s most of the phylogenetic analysis were performed only on the basis of sequence data. Quite surprisingly, despite the extremely low sequence identity, the PfkC family can be structurally classified as another member of the ribokinase group (Ito et al., 2001) which is now known as the ribokinase superfamily.

### 3. The ribokinase superfamily

Structurally, the PfkC and PfkB-like groups contain enzymes that present two domains. The large domain, which contains the core ribokinase-like fold, is an  $\alpha\beta\alpha$  structure where a central  $\beta$ -sheet mainly composed of parallel strands is flanked by  $\alpha$ -helices on both sides. Also, they present a smaller  $\beta$  domain which in general is used as a scaffold for dimerization (Sigrell et al., 1998). However, some of the enzymes are monomers. In this case, the hydrophobic core of the small domain is formed by the insertion of some  $\alpha$ -helices (Ito et al., 2001; Mathews et al., 1998). While not all the known PfkC enzymes are monomers (Jeong et al., 2003; Koga et al., 2000; Tuininga et al., 1999) all of them present those  $\alpha$ -helices in the small domain. Interestingly, the way in which many of them form multimers is not known, but seems to be highly enzyme specific.

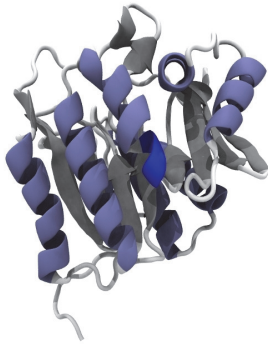
The active site of these enzymes is located in a cleft between both domains (Ito et al., 2001; Sigrell et al., 1998). For some members of the ribokinase family, it has been shown by means of x-ray crystallography that the relative orientation of the domains can be modified by the binding of the phosphoryl acceptor ligand (Schumacher et al., 2000; Sigrell et al., 1999) which has been suggested as a key step in the catalytic mechanism of these enzymes. In the PfkC case, a similar scenario has been suggested (Ito et al., 2003; Tsuge et al., 2002). Here, although the evidence is also crystallographic, it is indirect because the only enzyme crystallized in the apo form and complexed with a substrate is the ADP-dependent phosphofructokinase from *Pyrococcus horikoshii* (Currie et al., 2009) which does not show any domain movement. However, it was not possible to obtain a crystalline form of the enzyme in the presence of fructose-6-phosphate which could be the key component to induce the domain closing. In fact, it has been previously shown by us based on molecular modeling that the open conformation of these enzymes is most likely inactive (Merino & Guixé, 2008).

Table 1 shows most of the members of the ribokinase superfamily with known crystallographic structures. Based on this structural data it is possible to add other specificities



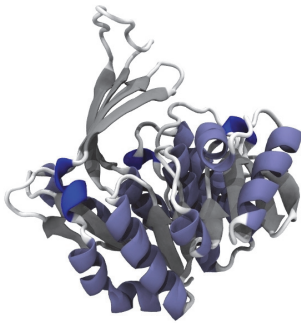
	PDB Code	Organism	Function
PfkC like	1UA4	<i>Pyrococcus furiosus</i>	Glucokinase
	1GC5	<i>Thermococcus litoralis</i>	Glucokinase
	1L2L	<i>Pyrococcus horikoshii</i>	Glucokinase
	1U2X	<i>Pyrococcus horikoshii</i>	Fructose-6-phosphate kinase
Vitamin kinase like	1JXH	<i>Salmonella typhimurium</i>	4-amino-5-hydroxymethyl-2-methylpyrimidine phosphate kinase
	1EKQ	<i>Bacillus subtilis</i>	Hydroxyethylthiazole kinase
	1V8A	<i>Pyrococcus horikoshii</i>	Hydroxyethylthiazole kinase
	1UB0	<i>Thermus thermophilus</i>	Phosphomethylpyrimidine kinase
	1LHP	<i>Ovis Aries</i>	Pyridoxal kinase
	1TD2	<i>Escherichia coli</i>	Pyridoxal kinase (PdxY)
	2DDM	<i>Escherichia coli</i>	Pyridoxal kinase (PdxK)
	2F7K	<i>Homo sapiens</i>	Pyridoxal kinase
	2I5B	<i>Bacillus subtilis</i>	Pyridoxal kinase
	1KYH	<i>Bacillus subtilis</i>	Unknown function
	2AX3	<i>Thermotoga maritima</i>	Unknown function
	2R3B	<i>Enterococcus faecalis</i>	Unknown function
PfkB like	2AFB	<i>Thermotoga maritima</i>	2-keto-3-deoxygluconate kinase
	2VAR	<i>Sulfolobus solfataricus</i>	2-keto-3-deoxygluconate kinase
	2DCN	<i>Sulfolobus tokodaii</i>	2-keto-3-deoxygluconate kinase
	1V1A	<i>Thermus thermophilus</i>	2-keto-3-deoxygluconate kinase
	2QCV	<i>Bacillus halodurans</i>	5-dehydro-2-deoxygluconate kinase
	1TZ6	<i>Salmonella enterica</i>	Aminoimidazol riboside kinase
	1BX4	<i>Homo sapiens</i>	Adenosine kinase
	1LII	<i>Toxoplasma gondii</i>	Adenosine kinase
	2PKN	<i>Mycobacterium tuberculosis</i>	Adenosine kinase
	2C49	<i>Methanocaldococcus jannaschii</i>	Nucleoside kinase
	1RKD	<i>Escherichia coli</i>	Ribokinase
	1VM7	<i>Thermotoga maritima</i>	Ribokinase
	2FV7	<i>Homo sapiens</i>	Ribokinase
	2QHP	<i>Bacteroides thetaiotaomicron</i>	Fructokinase
	2HW1	<i>Homo sapiens</i>	Ketohexokinase
	2ABQ	<i>Bacillus halodurans</i>	Fructose-1-phosphate kinase
	2F02	<i>Enterococcus Faecalis</i>	Tagatose-6-phosphate kinase
	2JG1	<i>Staphylococcus aureus</i>	Tagatose-6-phosphate kinase
	3CQD	<i>Escherichia coli</i>	Fructose-6-phosphate kinase
	3BF5	<i>Thermoplasma acidophilum</i>	Unknown function
	1VK4	<i>Thermotoga maritima</i>	Unknown function
	2NWH	<i>Agrobacterium tumefaciens</i>	Unknown function
	2RBC	<i>Agrobacterium tumefaciens</i>	Unknown function
	2AJR	<i>Thermotoga maritima</i>	Unknown function
	2JG5	<i>Staphylococcus aureus</i>	Unknown function

Table 1. Crystal structures of the ribokinase superfamily found in the PDB database.



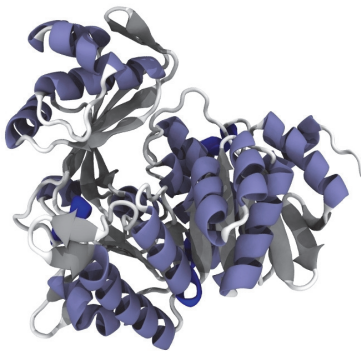
### **Vitamin kinase like branch**

Pyridoxal kinase  
Hydroxyethylthiazole kinase  
Phosphomethylpyrimidine kinase



### **PfkB like branch**

6-phosphofructokinase  
6-phosphotagatose-kinase  
1-phosphofructokinase  
Nucleoside kinase  
Adenosine kinase  
Ribokinase  
2-keto-3-deoxygluconate kinase  
Aminoimidazol riboside kinase



### **ADP-dependent branch (PfkC)**

Glucokinase  
6-phosphofructokinase

Fig. 1. Schematic representation of the three branches of the ribokinase superfamily. For the vitamin kinase like branch the pyridoxal kinase (pdxK) from *E. coli* (PDBID 2DDM) is used as example, for the PfkB like branch the ribokinase from *E. coli* (PDBID 1RKD) is used, and for the ADP-dependent branch the glucokinase from *T. litoralis* (PDBID 1GC5) is shown.

to the superfamily, such as adenosine kinase<sup>3</sup> (Mathews et al., 1998), 2-keto-3-deoxygluconate kinase (Ohshima et al., 2004), and aminoimidazole riboside kinase (Zhang et al., 2004). Beyond the sugar containing molecules, three dimensional structure comparison showed that kinases like 4-methyl-5- $\beta$ -hydroxyethylthiazole kinase (Campobasso et al., 2000), pyridoxal kinase (Li et al., 2002), and 4-amino-5-hydroxymethyl-2-methylpyrimidine phosphate kinase (Cheng et al., 2002) are also members of the ribokinase superfamily. Interestingly, these enzymes lack the small domain.

Already based on substrate specificities three major branches can be recognized (Figure 1). One of them contains those enzymes that catalyze the transfer of the  $\gamma$ -phosphate of ATP to molecules such as pyridoxal, or pyrimidine derivatives which we know as vitamin kinase like branch. The second contains all the enzymes that catalyze the transfer of the  $\gamma$ -phosphate of ATP to sugar containing molecules, such as fructose-6-phosphate, adenosine, aminoimidazole riboside, etc. We know this as the PfkB like branch. The last of them contains the enzymes that catalyze the transfer of the  $\beta$ -phosphate of ADP to glucose and fructose-6-phosphate which, as was mentioned before is known as PfkC family or ADP-dependent sugar kinase family. Based mainly on the presence of the small domain and the monomer complexity Zhang et al. (2004) proposed that the most ancient activity of the superfamily should be that catalyzed by the simplest enzyme which is 4-methyl-5- $\beta$ -hydroxyethylthiazole kinase. In that way, they propose that the increase of complexity in the monomers fold indicates a newer enzyme. By this hypothesis, the ADP-dependent enzymes and the monomeric adenosine kinases should be the newest acquisitions of the superfamily. However, this hypothesis was never tested. Nevertheless, although it could capture the essence of the evolutionary history of this group, considering the linearity of the hypothesis, it is rather unlikely that the true history of the group is entirely represented by it.

As it can be inferred from Figure 1 and Table 1 the ribokinase superfamily is an excellent example of how gene duplication has been used several times by nature to produce new specificities. This process has been recognized before as one of the most important steps in the creation of new protein functions (Chothia et al., 2003). Indeed, most of the proteins present inside a genome belong to a few protein families or a combination of them (see for example Chothia et al. (2003)). This degeneracy causes that the number of protein families represented in a genome are much smaller than the number of genes there.

Just as an example, a simple PSI-BLAST search on the genome of *E. coli* using the phosphofructokinase-2 as query finds 28 non-redundant proteins including: 6-phosphofructokinase, 1-phosphofructokinase, ribokinase, 2-keto-3-deoxygluconate kinase, and several proteins of unknown function. All of them present the PfkB-like fold (see Figure 1) which shows that this family is a very interesting example of gene duplications. However, the study of this feature is complicated by the lack of information on the function of several of the PfkB-like proteins.

#### 4. Structural evolution of the substrate specificity in the ADP-dependent sugar kinase family

The ADP-dependent sugar kinases have been found in several members of the *Pyrococcus*, *Thermococcus*, *Methanosarcina*, *Methanosaeta*, *Methanococcoides*, *Methanococcus*, *Methanocaldococcus*, and *Archaeoglobus* genera (Hansen & Schönheit, 2004; Kengen et al., 1994; Koga et al., 2000; Tuininga et al., 1999; Verhees et al., 2001). Also, it has been possible to

<sup>3</sup> These enzymes have a slightly different fold compared with the other nucleoside kinases (such as inosine-guanosine kinases) from the superfamily mentioned by Bork et al. (1993)

identify a distant homolog of these enzymes in the genome of higher eukaryotes, which has been proven to be an ADP-dependent glucokinase (Ronimus & Morgan, 2004). The metabolic role of the eukaryotic ADP-dependent glucokinases is unclear, but they have been suggested to be used in ischemic conditions (Ronimus & Morgan, 2004).

To date, the crystallographic structures of the ADP-dependent glucokinases from *Thermococcus litoralis* (Ito et al., 2001), *Pyrococcus horikoshii* (Tsuge et al., 2002), *Pyrococcus furiosus* (Ito et al., 2003), and the ADP-dependent phosphofructokinase from *Pyrococcus horikoshii* (Currie et al., 2009) have been solved. As opposed to the vitamin kinase or the PfkB-like branches of the ribokinase superfamily, to date just two specificities have been observed in the ADP-dependent branch (see Table 1). Considering that the ribokinase superfamily contains enzymes that catalyze the transfer the terminal phosphate of a nucleotide phosphate to the methyl alcohol end of a big number of small molecules which includes pyridoxal, pyrimidine derivatives, nucleosides, and several sugars, the PfkC family seems to be the one with the smallest substrate specificity in this group.

While there are many phosphoryl acceptor substrates in this superfamily, just two nucleotides, ADP and ATP, are described as the primary phosphoryl donors. Given the metabolic importance of the phosphoryl donor this specificity problem has received more attention than the acceptor problem in the literature. Of course, the specificity is not strict, and some other nucleotides can replace them. For instance, it has been shown that several ADP-dependent enzymes can use other purines (such as GDP) or even pyrimidines (such as UDP, (Currie et al., 2009)) as phosphoryl donors (Guixé & Merino, 2009). Also, GTP can be used by the phosphofructokinase-2 from *E. coli* and even produce substrate inhibition (unpublished results). Yet, it is important to remember that only those nucleotides with the right number of phosphates (either two for the ADP dependent enzymes or three for the ATP dependent) can be used, as it has been reviewed by us elsewhere (Guixé & Merino, 2009). This shows that any hint for the transition between nucleotide specificities has been obscured by evolution and specialization.

From an evolutionary perspective, while Zhang's hypothesis (Zhang et al., 2004) can be oversimplifying the problem, it captures the most common trend in the evolution of protein families: newer versions within the group tend to increase their structural complexity (Fong et al., 2007). Through their reasoning, ADP-dependent kinases should be closely related to the monomeric adenosine kinases. What has not been properly mentioned in the literature before is the fact that while the tertiary structure of the PfkC enzymes is almost equivalent to that of the PfkB enzymes, the topology of the C-terminal region is completely different (Figure 2). Indeed, this is the reason why it was not possible to group the ADP-dependent enzymes with the other members of the ribokinase superfamily just based on sequence comparison.

A BLAST search on the genome sequence of the archaeon *P. furiosus* reveals three PfkB-like enzymes of unknown function. As it is also possible to find vitamin kinase like enzymes in the genome of the *thermococcales* (see for instance Table 1), then it is possible to deduce that all three modern branches of the ribokinase superfamily have been originated by ancient gene duplication events followed by extensive topological modifications. While the addition of the small domain can be viewed as a trivial modification since it only involves the insertion of sequence, the C-terminal topological reordering involves a non-cyclic permutation. Now, considering that the ADP-dependent enzymes should be the modern ones, in order to be compatible with the topological reordering an ATP-dependent enzyme should present an extra strand in the C-terminal end of the protein extending the central  $\beta$ -sheet. Figure 2 shows that indeed, this requirement is fulfilled by some PfkB-like enzymes. Quite surprisingly, the sugar-phosphate kinases and not adenosine kinases are those who show the extra strand. This

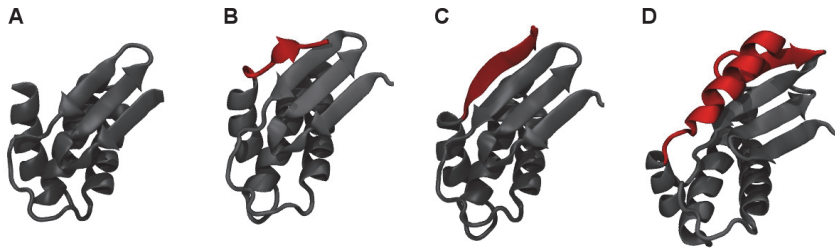


Fig. 2.  $\beta$ -meander region of several members of the ribokinase superfamily such as Pfk-2 from *E. coli* (A), a fructose-1-phosphate kinase from *B. halodurans* (B), a putative phosphofructokinase from *S. aureus* (C), and the ADP-dependent glucokinase from *P. furiosus* (D). In red is shown the C-terminal extension thought to be needed for the circular permutation.

suggests that the ADP-dependent enzymes are most closely related to the other glycolytic enzymes present in the superfamily, which seems to be reasonable given the similarity of their substrates.

Indeed, this C-terminal reordering is quite interesting since this same region constitutes almost all the nucleotide binding site. However, while this permutation almost certainly alters the dynamics of the binding pocket, we are not sure if it will alter the specificity of the enzyme. Nevertheless, it requires empirical testing which is now being performed in our laboratory.

Interestingly, the  $\alpha$  and  $\beta$  phosphates of ADP are accommodated in the binding site of the PfkC enzymes almost in the same way as the  $\beta$  and  $\gamma$  phosphates of ATP in the remaining members of the superfamily. This led Ito et al. (2001) to suggest that the bulky side chain of Y357 in the ADP-dependent glucokinase from *T. litoralis* which is located below the ribose moiety of ADP was pushing the nucleotide forward and then rendering an enzyme unable to use ATP. However, we indirectly demonstrated that this is not the case since for the ADP-dependent phosphofructokinase from *P. horikoshii* the presence of a significantly less bulky side chain (I340) does not produce an enzyme with ATP-dependent activity (Currie et al., 2009).

While in most of the members of the *Euryarchaeota* there are two ADP-dependent enzymes coded in their genomes, the archaeon *M. jannaschii* presents just one copy of these genes. Surprisingly, the enzyme is able to catalyze the transfer of the  $\beta$ -phosphate of ADP to either glucose or fructose-6-phosphate (Sakuraba et al., 2002). Based on this feature, it was proposed that this enzyme represent an ancestral state of the family, which later gave rise to the separate specificities through a gene duplication event (Sakuraba et al., 2002). However, this hypothesis had to wait six years to be tested (Merino & Guixé, 2008).

To test this hypothesis we used the Bayesian method of phylogenetic inference implemented in the MrBayes 3.1 software (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003). Initially, a structural based sequence profile was built by means of a structural alignment of the ADP-dependent glucokinases from *T. litoralis*, *P. horikoshii*, and *P. furiosus* and the ADP-dependent phosphofructokinase from *P. horikoshii*. Later all the ADP-dependent kinases from archaeal source were aligned to this profile. After several rounds of alignment refinement the eukaryotic ADP-dependent enzymes were added. As they share only about 15 to 20% sequence identity with the archaeal versions the alignment was guided by secondary structure predictions.



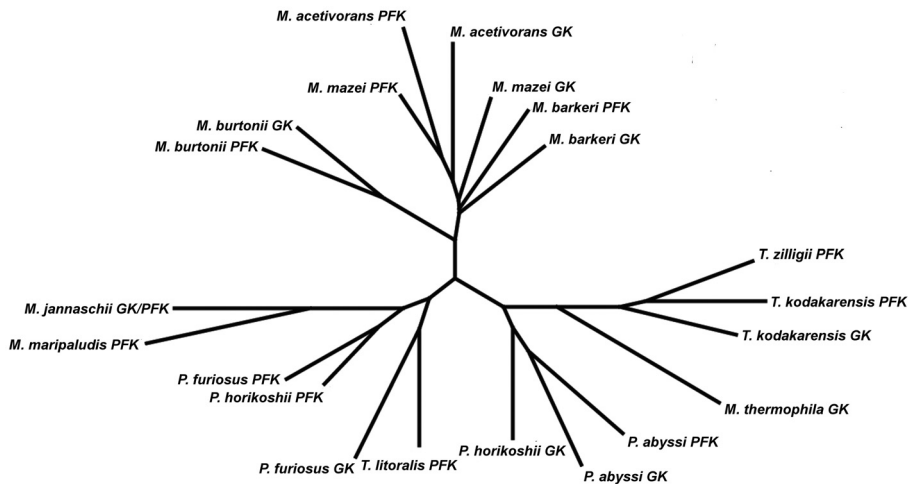


Fig. 4. Dendrogram grouping the archaeal ADP-dependent genes according to their average difference in relative synonymous codon usage. Modified from (Merino & Guixé, 2008).

*methanosarcinales*. Indeed a similar scenario for the generation of paralogous genes has been proposed before (Gogarten et al., 2002).

To test this hypothesis we analyzed the relative synonymous codon usage of the archaeal PfkC genes (McInerney, 1998). By this methodology, the frequency of any given codon in a gene is calculated relative to the frequency expected for an unbiased codon usage. Figure 4 shows that, in general, genes are grouped very close to their paralogous. If this is not the case, they are at least inside a group that contains closely related species. The only exception is the glucokinase from *Methanosaeta thermophila* which is located inside the *thermococcales* group (Figure 4). Indeed, when the codon usage of this gene is compared with the codon usage of the archaeal genomes, it seems to be more related to the genome of *T. litoralis* than to its own genome (not shown).

While the data present above are not enough to prove the horizontal transfer hypothesis it still strongly suggests that this process has been involved in the evolution of the ADP-dependent sugar kinase family. It is important to stress out that if the event of horizontal gene transfer is ancient enough, then the accumulation of a sufficient number of mutations should have masked it. If this is our case then, to our knowledge, there is no sequence based technique to prove the hypothesis.

Sakuraba et al. (2002) demonstrated that when the bifunctional enzyme was using fructose-6-phosphate as substrate glucose can act as a competitive inhibitor. They proposed that this was produced because both sugars bind to the same site. It is important to mention that competitive inhibition does not necessarily indicates that substrate and inhibitor have the same site, but in this case it is certainly the case. To take advantage of this fact we modeled the bifunctional enzyme and its interaction with both sugars. In this way, it is possible to gain as much information as possible about the structural determinants of the sugar specificity.

Figure 5 shows the predicted interaction geometries for both substrates. For clarity just the residues in a 5 Å radius are shown. As it was inferred by Sakuraba et al. (2002) the interaction between the protein and both substrates are very similar. Indeed, just three of the residues seems to differ significantly in the way they interact with the sugars. For instance, while



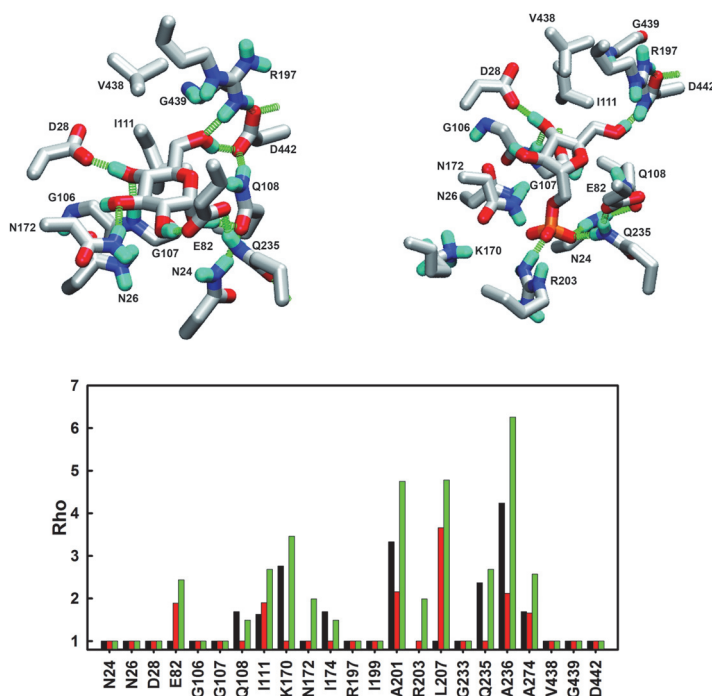


Fig. 5. **Left.** Glucose docked to the bifunctional enzyme. **Right.** Fructose-6-phosphate docked to the bifunctional enzyme. **Bottom.** Results of the real value evolutionary trace analysis for all the residues within 5 Å from the ligands. The results for the glucokinase specificity are shown in black, those for the phosphofructokinase specificity in red, and those for the whole family in green. Modified from (Merino & Guixé, 2008).

E82 makes a hydrogen bond with the hydroxyl located at C2 in glucose it does not seem to interact in any specific way with fructose-6-phosphate. Indeed, this side chain has been proposed by other authors as key for the glucokinase specificity (Ito et al., 2003; Sakuraba et al., 2002). On the other hand, R203 is making a close salt bridge with the phosphate moiety of fructose-6-phosphate while it does not interact with glucose. Although K170 is not in the 5 Å radius we had strong evidence that, as in the R203 case, this side chain was also involved in the phosphate binding (see below).

To quantify the conservation degree of the residues inside the sugar binding site we used a tree-based residue ranking system called real value evolutionary trace (Mihalek et al., 2004). Briefly, the method ranks the residues as follows:

First, let us consider a rooted evolutionary tree with  $N$  leaves (sequences). If we number the nodes in the tree starting with the root being 1 then, using as example Figure 3, the node number 2 should be the one with 0.98 posterior probability, and so on. Using this method it is possible to number  $N - 1$  nodes. Each node defines some groups  $g$  of sequences. The root node of course creates a group with all of them. Node number 2 creates a group that contains the ADP-dependent glucokinases from *methanosarcinales* and other with the rest of



the sequences. By this nomenclature one can define a measurement of the conservation of each position in the alignment  $i$  where

$$r_i = 1 + \sum_{n=1}^{N-1} \begin{cases} 0 & \text{if position } i \text{ conserved within each group } g \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

It is clear that if a residue is conserved from the root of the tree, then it will have a  $r_i$  of 1. As it gets less and less conserved  $r_i$  will be higher. To account for the sequence conservation within each group, the  $r_i$  value was weighted by sequence entropy given the expression

$$\rho_i = 1 + \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^n \left( - \sum_{aa=1}^{20} f_{ia}^g \log f_{ia}^g \right) \quad (2)$$

where  $f_{ia}^g$  stands for the frequency of appearance of amino acid  $a$  inside the group  $g$ .

Figure 5 (bottom) shows the result of the ranking applied to the whole PfkC family, and both separated specificities. It is clear from the figure that most of the residues are conserved in the whole family. Interestingly, E82 is only conserved inside the glucokinase specificity, which is in good agreement with the role proposed above. Also, K170 and R203 are only conserved inside the phosphofructokinase specificity which makes them the inverse case of the E82 residue.

Interestingly, N172 is conserved inside both specificities, but it is not in the whole family. The reason for this is that within phosphofructokinases this residue is strictly an asparagine while inside the glucokinases is always a histidine. This suggests that this residue is also related with sugar specificity, but the reason is not as clear as the above examples.

Recently, we used a more elegant method known as explicit likelihood of subset covariation (ELCS) (Dekker et al., 2004) to explore the correlation between mutations to search for the structural specificity determinants.

Figure 6 shows the group of side chains with the highest ELCS score. Surprisingly, the group contain a side chain that belongs to a highly conserved motif called NXXE which has been related with metal binding to the enzymes of the superfamily (Maj et al., 2002; Parducci et al., 2006; Rivas-Pardo et al., 2011). In fact, we have demonstrated that this motif is related to the binding of the catalytic and regulatory metals in the ADP-dependent sugar kinase family (To be published). Also, the first group found by the ELCS method contains some residues that we proposed before as specificity related. The role of the R48/D65<sup>4</sup>, R65/S76, P73/F90 mutations is not clear, but seem to be related to the dynamics of the small domain. K158/C174 (equivalent to K170 in the bifunctional enzyme), N160/H176 (equivalent to N172 in the bifunctional enzyme), and R191/D203 (equivalent to R203 in the bifunctional enzyme) are clearly interacting with the sugars. Interestingly, when the position R191/D203 presents an arginine, this positive side chain coordinates the phosphate group present in the fructose-6-phosphate molecule. On the other hand, when it presents an aspartic acid, this side chain interacts with the histidine in the N160/H176 position, allowing the histidine to be correctly positioned to make an h-bond with the O2 hydroxyl group of glucose. Curiously, the position equivalent to E82 from the bifunctional enzyme does not appear to be correlated with other positions by the ELCS method. However this could be due to the small amount of sequence information used for the analysis.

<sup>4</sup> We use the numbering of *Ph*PFK/*Pf*GK for the correlated mutations. See Figure 6 for clarity.

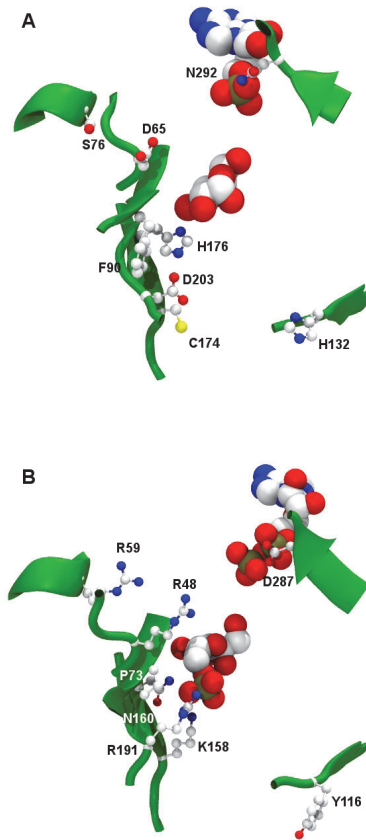


Fig. 6. First cluster of correlated mutations in the PfkC family identified by the ELCS method. **A.** Crystal structure of the glucokinase from *P. furiosus*. Glucose and AMP are shown. **B.** Structural model for the ternary complex between the phosphofructokinase from *P. horikoshii*, ADP and fructose-6-phosphate. The coordinates were derived from the molecular dynamics simulation performed in (Currie et al., 2009)

We have tested the predictions made by the evolutionary trace and the ELCS methods by means of mutagenesis using the ADP-dependent phosphofructokinase from *horikoshii* (Currie et al., 2009) as a model.

Table 2 shows the effect of each mutation on the kinetic parameters of fructose-6-phosphate and MgADP. As it can be predicted, either the mutations R191A, R191E, or K158A produce a high increase in the  $K_M$  value for fructose-6-phosphate with a little effect on  $k_{cat}$  or  $K_M$  from MgADP.

On the other hand, the N160A increase three-fold  $k_{cat}$  with a concomitant high increase in the  $K_M$  value for fructose-6-phosphate. The reason for the increase in activity is not clear, but it suggests that while this interaction increases the affinity of the protein for the sugar it

Enzyme	$k_{cat}$	$K_M$ F6P	$k_{cat}/K_M$ F6P
	$s^{-1}$	$\mu M$	$M^{-1}s^{-1}$
<b>Wild Type</b>	$45.5 \pm 4.0$	$15.2 \pm 2.5$	$2.98 \cdot 10^6$
<b>A71E</b>	$39.3 \pm 9.9$	$22.2 \pm 2.6$	$1.77 \cdot 10^6$
<b>K158A</b>	$41.0 \pm 6.7$	$6500 \pm 1300$	$6.30 \cdot 10^3$
<b>N160A</b>	$151 \pm 16$	$415 \pm 13$	$3.65 \cdot 10^5$
<b>N160Q</b>	$14.7 \pm 0.6$	$6300 \pm 720$	$2.33 \cdot 10^3$
<b>R191A</b>	$27.4 \pm 1.5$	$254.4 \pm 26.1$	$1.07 \cdot 10^5$
<b>R191E</b>	$42.5 \pm 1$	$4870 \pm 170$	$8.73 \cdot 10^3$

Table 2. Kinetic parameters of wild type and mutant versions of the ADP-dependent phosphofructokinase from *P. horikoshii*. Modified from (Currie et al., 2009). All experiments were performed at 50 °C.

imposes a strain in the transition state, which results in a decrease of  $k_{cat}$ . However, none of these mutations produce an enzyme with glucokinase activity.

The A71E mutation does not affect the catalytic constants nor  $K_M$  for fructose-6-phosphate. Surprisingly, it produces an enzymes that now can catalyze the transfer of the  $\beta$ -phosphate of ADP to glucose with a  $k_{cat}$  of  $2.7 \pm 0.05 s^{-1}$  and a  $K_M$  of  $3.95 \pm 0.2 mM^5$ . Also we recently have produced a N160H mutant. It dramatically increases the  $K_M$  value for fructose-6-phosphate to  $6.3 \pm 0.72 mM$  and decreases the  $k_{cat}$  almost four-fold (Table 2). As in the A71E case, it also produces a bifunctional enzyme that can use glucose as substrate. However, for this mutant no clear saturation is seen even for 25 mM glucose. Based on a Lineweaver-Burk plot, it is possible to estimate a  $k_{cat}$  of  $2.42 s^{-1}$  and a  $K_M$  value of 25.3 mM. Clearly, this mutation produces a much stronger effect than A71E.

The last two mutations are key to understanding the specialization problem since they not only enable the phosphofructokinase from *P. horikoshii* to use glucose as substrate. Competition experiments with this enzyme have shown that glucose does not bind to the wild type version, which demonstrates that the mutations somehow unblock the binding site for the binding of glucose. Curiously, both mutations points to the interaction between the protein and the hydroxyl group at C2 of glucose. This suggests that the specificity determinants are not evenly distributed amongst the binding site, but rather concentrated in hot-spots. In this light, in order to revert the specificities, just a couple of mutations are needed.

## 5. Are two ADP-dependent kinases better than one?

Considering that the glycolysis of *M. jannaschii* is functional with just one enzyme in charge of the phosphorylation of glucose and fructose-6-phosphate it is not clear, at a first glance, why two genes were select by nature in the other members of the *Euryarchaeota*. As it was mentioned above, glucokinases are on the top of several pathways and hence the modification of their activity affects a big part of the metabolism. Indeed, this enzyme generally have a great control of the carbon flux. On the other hand, phosphofructokinases seem to be closely related with the balance between glycolysis and gluconeogenesis. In the archaeon *P. furiosus* it has been shown that the switching between these two metabolic pathways is controlled at the expression level (Schut et al., 2003). When the ADP-dependent phosphofructokinase is expressed the fructose-1,6-bisPase is repressed and *vice versa*. Of course, just shutting the

<sup>5</sup> Glucokinase experiments were performed at 40 °C given the instability of the auxiliar enzyme used.

bifunctional enzyme down in *M. jannaschii* will not only decrease the phosphofructokinase activity, but it will have the undesirable side effect of decreasing the glucokinase activity. In this light, the use of one enzyme for each specificity has a great impact in how the cell can regulate the carbon flux. Indeed, the fact the sugar specificity residues are correlated with some others related with regulation (such as the mutation in the NXXE motif) strongly favors our explanation.

What kind of process produce this? It is now a generally accepted hypothesis that less important genes or parts of a gene tend to change or evolve faster than less important ones, which is known as the Kimura-Ohta principle (for a review see Camps et al. (2007)). It is clear that the upon a gene duplication, the phenotypic effect of a mutation in any copy of these genes should be fairly null with the only exception of those that produces specialized genes. It can be argued that even those mutations that produce inactive proteins which should be deleterious and removed by purifying selection under normal conditions are now nearly neutral since the only extra cost is to produce a non-functional protein.

It has been shown recently that upon a change in the fitness optimum (either produced by an environmental change or an internal redistribution of fluxes) most mutations are fixed by natural selection up until the genes reach a nearly optimal sequence. Then they accumulated mutations according to a neutral model (Razeto-Barry et al., 2011). From the arguments given above, it is clear that the only way in which a duplicated gene can break the neutral regime is when a rare specializing mutation is fixed. In this case, the organism must adapt to the new distribution of internal fluxes. Ohta (1987) reached a similar conclusion based on simulations. He stated: "Positive natural selection favors those chromosomes with more beneficial mutations in redundant copies than others in the population, but accumulation of deteriorating mutations (pseudo genes) have no effect on fitness so long as there remains a functional gene. The results imply the following: Positive natural selection is needed in order to acquire gene families with new functions. Without it, too many pseudo genes accumulate before attaining a functional gene family".

As we have shown here, for our protein family, this would imply just one or two mutations since for instance, just the change of a single interaction can change the balance between both specificities. Of course, upon specialization, mutations that modulate regulation (such as those related with the NXXE motif) increases the flexibility of the metabolism.

Interestingly, although most of the *Euryarchaeota* that present the ADP-dependent kinases have two separated specificities the glucokinase from psychrophilic archaeon *Methanococcoides burtonii* have a big C-terminal deletion that should make it non-functional (Merino & Guixé, 2008). The fact that it is still possible to know that it was a glucokinase suggests that this deletion was recent. The phosphofructokinase gene in this archaeon present a glutamic acid in the position equivalent to E82 in the bifunctional enzyme, which suggests that this could be a bifunctional enzyme too. In this way, it appears that until the phosphofructokinase gene is entirely specialized it still exist the possibility of losing the specific glucokinase gene.

## 6. Concluding remarks

Our studies about the evolution of the ADP-dependent sugar kinase family showed that the root of the family is located inside the glucokinase group, demonstrating that the bifunctional (glucokinase/phosphofructokinase) enzyme is not an ancestral form, but could be a transitional form from glucokinase to phosphofructokinase. Unfortunately, to date it has not been possible to obtain the crystal structure of any ADP-dependent phosphofructokinase in the presence of fructose-6-phosphate. However, based on structural modeling we have been able to understand partially the structure/specificity relation up to the point where we

can produce bifunctional enzymes from specific ones. Strikingly, the sugar discrimination is somehow concentrated in very few hot-spots in the structure. Indeed, the introduction of just one hydrogen bond or some salt bridges seems to modulate the affinity for glucose or fructose-6-phosphate respectively. Unfortunately, to date, we have been unable to absolutely switch the specificity of these enzymes.

The gene duplication event itself seems to be related with the separated regulation of the glucokinase and phosphofructokinase activity, along with the balance between glycolysis and gluconeogenesis. Indeed, with two different enzymes a finest tuning of the carbon flux inside the cell can be achieved.

## 7. Acknowledgements

We are very grateful to Dr. Ricardo Cabrera for his help in the creation of this chapter by means of enlightening discussions about the evolutionary implications of the gene duplication in the modified Embden-Meyerhof pathway. This work was supported by Fondo Nacional de Desarrollo Científico y Tecnológico (Fondecyt) through the grant 1110137.

## 8. References

- Allers, T. & Mevarech, M. (2005). Archaeal genetics - the third way., *Nat. Rev. Genet.* 6(1): 58–73.
- Atkinson, D. E. & Walton, G. M. (1965). Kinetics of regulatory enzymes. *Escherichia coli* phosphofructokinase., *J. Biol. Chem.* 240: 757–763.
- Bork, P., Sander, C. & Valencia, A. (1993). Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases., *Protein Sci.* 2(1): 31–40.
- Cabrera, R., Babul, J. & Guixé, V. (2010). Ribokinase family evolution and the role of conserved residues at the active site of the PfkB subfamily representative, Pfk-2 from *Escherichia coli*., *Arch. Biochem. Biophys.* 502(1): 23–30.
- Campobasso, N., Mathews, I. I., Begley, T. P. & Ealick, S. E. (2000). Crystal structure of 4-methyl-5-beta-hydroxyethylthiazole kinase from *Bacillus subtilis* at 1.5 Å resolution., *Biochemistry* 39(27): 7868–7877.
- Camps, M., Herman, A., Loh, E. & Loeb, L. A. (2007). Genetic constraints on protein evolution., *Crit. Rev. Biochem. Mol. Biol.* 42(5): 313–326.
- Cheng, G., Bennett, E. M., Begley, T. P. & Ealick, S. E. (2002). Crystal structure of 4-amino-5-hydroxymethyl-2-methylpyrimidine phosphate kinase from *Salmonella typhimurium* at 2.3 Å resolution., *Structure* 10(2): 225–235.
- Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. (2003). Evolution of the protein repertoire., *Science* 300(5626): 1701–1703.
- Currie, M. A., Merino, F., Skarina, T., Wong, A. H. Y., Singer, A., Brown, G., Savchenko, A., Caniuguir, A., Guixé, V., Yakunin, A. F. & Jia, Z. (2009). ADP-dependent 6-phosphofructokinase from *Pyrococcus horikoshii* OT3: structure determination and biochemical characterization of PH1645., *J. Biol. Chem.* 284(34): 22664–22671.
- Dekker, J. P., Fodor, A., Aldrich, R. W. & Yellen, G. (2004). A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments., *Bioinformatics* 20(10): 1565–1572.
- Dörr, C., Zaparty, M., Tjaden, B., Brinkmann, H. & Siebers, B. (2003). The hexokinase of the hyperthermophile *Thermoproteus tenax*. ATP-dependent hexokinases and ADP-dependent glucokinases, two alternatives for glucose phosphorylation in Archaea., *J. Biol. Chem.* 278(21): 18744–18753.

- Evans, P. R., Farrants, G. W. & Hudson, P. J. (1981). Phosphofructokinase: structure and control., *Philos. Trans. R Soc. Lond. B Biol. Sci.* 293(1063): 53–62.
- Fong, J. H., Geer, L. Y., Panchenko, A. R. & Bryant, S. H. (2007). Modeling the evolution of protein domain architectures using maximum parsimony., *J. Mol. Biol.* 366(1): 307–315.
- Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer., *Mol. Biol. Evol.* 19(12): 2226–2238.
- Guixé, V. & Merino, F. (2009). The ADP-dependent sugar kinase family: kinetic and evolutionary aspects., *IUBMB Life* 61(7): 753–761.
- Hansen, T. & Schönheit, P. (2004). ADP-dependent 6-phosphofructokinase, an extremely thermophilic, non-allosteric enzyme from the hyperthermophilic, sulfate-reducing archaeon *Archaeoglobus fulgidus* strain 7324., *Extremophiles* 8(1): 29–35.
- Huelsenbeck, J. P. & Ronquist, F. (2001). MrBayes: Bayesian inference of phylogenetic trees., *Bioinformatics* 17(8): 754–755.
- Ito, S., Fushinobu, S., Jeong, J.-J., Yoshioka, I., Koga, S., Shoun, H. & Wakagi, T. (2003). Crystal structure of an ADP-dependent glucokinase from *Pyrococcus furiosus*: implications for a sugar-induced conformational change in ADP-dependent kinase., *J. Mol. Biol.* 331(4): 871–883.
- Ito, S., Fushinobu, S., Yoshioka, I., Koga, S., Matsuzawa, H. & Wakagi, T. (2001). Structural basis for the ADP-specificity of a novel glucokinase from a hyperthermophilic archaeon., *Structure* 9(3): 205–214.
- Jeong, J.-J., Fushinobu, S., Ito, S., Shoun, H. & Wakagi, T. (2003). Archaeal ADP-dependent phosphofructokinase: expression, purification, crystallization and preliminary crystallographic analysis., *Acta Crystallogr. D Biol. Crystallogr.* 59(Pt 7): 1327–1329.
- Jones, W., Leigh, J., Mayer, F., Woese, C. & Wolfe, R. (1983). *Methanococcus jannaschii* sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent, *Arch. Microbiol.* 136(4): 254–261.
- Kengen, S. W., de Bok, F. A., van Loo, N. D., Dijkema, C., Stams, A. J. & de Vos, W. M. (1994). Evidence for the operation of a novel embden-meyerhof pathway that involves ADP-dependent kinases during sugar fermentation by *Pyrococcus furiosus*., *J Biol Chem* 269(26): 17537–17541.
- Koga, S., Yoshioka, I., Sakuraba, H., Takahashi, M., Sakasegawa, S., Shimizu, S. & Ohshima, T. (2000). Biochemical characterization, cloning, and sequencing of ADP-dependent (AMP-forming) glucokinase from two hyperthermophilic archaea, *Pyrococcus furiosus* and *Thermococcus litoralis*., *J. Biochem.* 128(6): 1079–1085.
- Kotlarz, D. & Buc, H. (1981). Regulatory properties of phosphofructokinase 2 from *Escherichia coli*., *Eur. J. Biochem.* 117(3): 569–574.
- Li, M., Kwok, F., Chang, W., Lau, C., Zhang, J., Lo, S. C. L., Jiang, T. & Liang, D. (2002). Crystal structure of brain pyridoxal kinase, a novel member of the ribokinase superfamily., *J. Biol. Chem.* 277(48): 46385–46390.
- Maj, M. C., Singh, B. & Gupta, R. S. (2002). Pentavalent ions dependency is a conserved property of adenosine kinase from diverse sources: identification of a novel motif implicated in phosphate and magnesium ion binding and substrate inhibition., *Biochemistry* 41(12): 4059–4069.
- Mathews, I. I., Erion, M. D. & Ealick, S. E. (1998). Structure of human adenosine kinase at 1.5 Å resolution., *Biochemistry* 37(45): 15607–15620.
- McInerney, J. O. (1998). GCUA: general codon usage analysis., *Bioinformatics* 14(4): 372–373.

- Merino, F. & Guixé, V. (2008). Specificity evolution of the ADP-dependent sugar kinase family: *in silico* studies of the glucokinase/phosphofructokinase bifunctional enzyme from *Methanocaldococcus jannaschii*, *FEBS J.* 275(16): 4033–4044.
- Mihalek, I., Res, I. & Lichtarge, O. (2004). A family of evolution-entropy hybrid methods for ranking protein residues by importance., *J. Mol. Biol.* 336(5): 1265–1282.
- Mukund, S. & Adams, M. W. (1991). The novel tungsten-iron-sulfur protein of the hyperthermophilic archaeobacterium, *Pyrococcus furiosus*, is an aldehyde ferredoxin oxidoreductase. evidence for its participation in a unique glycolytic pathway., *J Biol Chem* 266(22): 14208–14216.
- Mukund, S. & Adams, M. W. (1995). Glyceraldehyde-3-phosphate ferredoxin oxidoreductase, a novel tungsten-containing enzyme with a potential glycolytic role in the hyperthermophilic archaeon *Pyrococcus furiosus*., *J. Biol. Chem.* 270(15): 8389–8392.
- Ohshima, N., Inagaki, E., Yasuike, K., Takio, K. & Tahirov, T. H. (2004). Structure of *Thermus thermophilus* 2-keto-3-deoxygluconate kinase: evidence for recognition of an open chain substrate., *J. Mol. Biol.* 340(3): 477–489.
- Ohta, T. (1987). Simulating evolution by gene duplication., *Genetics* 115(1): 207–213.
- Parducci, R. E., Cabrera, R., Baez, M. & Guixé, V. (2006). Evidence for a catalytic  $Mg^{2+}$  ion and effect of phosphate on the activity of *Escherichia coli* phosphofructokinase-2: regulatory properties of a ribokinase family member., *Biochemistry* 45(30): 9291–9299.
- Peng, Z. Y. & Mansour, T. E. (1992). Purification and properties of a pyrophosphate-dependent phosphofructokinase from *Toxoplasma gondii*., *Mol. Biochem. Parasitol.* 54(2): 223–230.
- Razeto-Barry, P., Díaz, J., Cotoras, D. & Vásquez, R. A. (2011). Molecular evolution, mutation size and gene pleiotropy: a geometric reexamination., *Genetics* 187(3): 877–885.
- Rivas-Pardo, J. A., Caniuguir, A., Wilson, C. A. M., Babul, J. & Guixé, V. (2011). Divalent metal cation requirements of phosphofructokinase-2 from *E. coli*. evidence for a high affinity binding site for  $Mn^{2+}$ ., *Arch. Biochem. Biophys.* 505(1): 60–66.
- Ronimus, R. S. & Morgan, H. W. (2004). Cloning and biochemical characterization of a novel mouse ADP-dependent glucokinase., *Biochem. Biophys. Res. Commun.* 315(3): 652–658.
- Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models., *Bioinformatics* 19(12): 1572–1574.
- Sakuraba, H., Goda, S. & Ohshima, T. (2004). Unique sugar metabolism and novel enzymes of hyperthermophilic archaea., *Chem. Rec.* 3(5): 281–287.
- Sakuraba, H., Yoshioka, I., Koga, S., Takahashi, M., Kitahama, Y., Satomura, T., Kawakami, R. & Ohshima, T. (2002). ADP-dependent glucokinase/phosphofructokinase, a novel bifunctional enzyme from the hyperthermophilic archaeon *Methanococcus jannaschii*., *J. Biol. Chem.* 277(15): 12495–12498.
- Sapra, R., Bagramyan, K. & Adams, M. W. W. (2003). A simple energy-conserving system: proton reduction coupled to proton translocation., *Proc. Natl. Acad. Sci. USA* 100(13): 7545–7550.
- Schirmer, T. & Evans, P. (1990). Structural basis of the allosteric behaviour of phosphofructokinase, *Nature* 343: 140–145.
- Schumacher, M. A., Scott, D. M., Mathews, I. I., Ealick, S. E., Roos, D. S., Ullman, B. & Brennan, R. G. (2000). Crystal structures of *Toxoplasma gondii* adenosine kinase reveal a novel catalytic mechanism and prodrug binding., *J. Mol. Biol.* 298(5): 875–893.
- Schut, G. J., Brehm, S. D., Datta, S. & Adams, M. W. W. (2003). Whole-genome dna microarray analysis of a hyperthermophile and an archaeon: *Pyrococcus furiosus* grown on carbohydrates or peptides., *J. Bacteriol.* 185(13): 3935–3947.



- Sigrell, J. A., Cameron, A. D., Jones, T. A. & Mowbray, S. L. (1998). Structure of *Escherichia coli* ribokinase in complex with ribose and dinucleotide determined to 1.8 Å resolution: insights into a new family of kinase structures., *Structure* 6(2): 183–193.
- Sigrell, J. A., Cameron, A. D. & Mowbray, S. L. (1999). Induced fit on sugar binding activates ribokinase., *J. Mol. Biol.* 290(5): 1009–1018.
- Torres, J. C., Guixé, V. & Babul, J. (1997). A mutant phosphofructokinase produces a futile cycle during gluconeogenesis in *Escherichia coli*., *Biochem. J.* 327: 675–684.
- Torres, N. V., Mateo, F. & Meléndez-Hevia, E. (1988). Shift in rat liver glycolysis control from fed to starved conditions. flux control coefficients of glucokinase and phosphofructokinase., *FEBS Lett.* 233(1): 83–86.
- Tsuge, H., Sakuraba, H., Kobe, T., Kujime, A., Katunuma, N. & Ohshima, T. (2002). Crystal structure of the ADP-dependent glucokinase from *Pyrococcus horikoshii* at 2.0-Å resolution: a large conformational change in ADP-dependent glucokinase., *Protein Sci.* 11(10): 2456–2463.
- Tuininga, J. E., Verhees, C. H., van der Oost, J., Kengen, S. W., Stams, A. J. & de Vos, W. M. (1999). Molecular and biochemical characterization of the ADP-dependent phosphofructokinase from the hyperthermophilic archaeon *Pyrococcus furiosus*., *J. Biol. Chem.* 274(30): 21023–21028.
- van Rooijen, R. J., van Schalkwijk, S. & de Vos, W. M. (1991). Molecular cloning, characterization, and nucleotide sequence of the tagatose 6-phosphate pathway gene cluster of the lactose operon of *Lactococcus lactis*., *J. Biol. Chem.* 266(11): 7176–7181.
- Verhees, C. H., Kengen, S. W. M., Tuininga, J. E., Schut, G. J., Adams, M. W. W., De Vos, W. M. & Van Der Oost, J. (2003). The unique features of glycolytic pathways in Archaea., *Biochem. J.* 375(2): 231–246.
- Verhees, C. H., Tuininga, J. E., Kengen, S. W., Stams, A. J., van der Oost, J. & de Vos, W. M. (2001). ADP-dependent phosphofructokinases in mesophilic and thermophilic methanogenic archaea., *J. Bacteriol.* 183(24): 7145–7153.
- Woese, C. R. & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms., *Proc. Natl. Acad. Sci. USA* 74(11): 5088–5090.
- Woese, C. R., Kandler, O. & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya., *Proc. Natl. Acad. Sci. USA* 87(12): 4576–4579.
- Wu, L. F., Reizer, A., Reizer, J., Cai, B., Tomich, J. M. & Saier, M. H. (1991). Nucleotide sequence of the *Rhodobacter capsulatus* fruK gene, which encodes fructose-1-phosphate kinase: evidence for a kinase superfamily including both phosphofructokinases of *Escherichia coli*., *J. Bacteriol.* 173(10): 3117–3127.
- Zhang, Y., Dougherty, M., Downs, D. M. & Ealick, S. E. (2004). Crystal structure of an aminoimidazole riboside kinase from *Salmonella enterica*: implications for the evolution of the ribokinase superfamily., *Structure* 12(10): 1809–1821.



# Duplication of Coagulation Factor Genes and Evolution of Snake Venom Prothrombin Activators

Shiyang Kwong<sup>1</sup> and R. Manjunatha Kini<sup>1,2</sup>

<sup>1</sup>*Department of Biological Sciences, Faculty of Science  
National University of Singapore, Singapore*

<sup>2</sup>*Department of Biochemistry, Medical College of Virginia  
Virginia Commonwealth University, Richmond, Virginia*

<sup>1</sup>*Singapore*

<sup>2</sup>*USA*

## 1. Introduction

Snake venom is a complex mixture of pharmacologically active molecules which are responsible for immobilization, paralysis, death and digestion of prey organisms. This armory of toxins has evolved to target two key systems, namely the neuromuscular and circulatory systems, in order to induce rapid immobilization and death. So far, several hundreds of protein toxins from snake venoms have been purified and characterized. Most of these toxins have been documented to be structurally, and at times functionally, similar to proteins expressed in different tissues of the body. For example, elapid phospholipase A<sub>2</sub> toxins are structurally and catalytically similar to mammalian pancreatic phospholipase A<sub>2</sub> enzymes (Robin Doley et al. 2009). Similarly, sarafotoxins are structurally and functionally similar to endothelins produced primarily in endothelium (Landan et al. 1991a). Based on such structural and functional similarities, it is hypothesized that toxin proteins are “recruited” from body proteins by gene duplication (Fry 2005). Accordingly, the genes of body proteins are duplicated and modified to have differential and specific expression in venom glands. This phenomenon is broadly termed as “recruitment”. This “recruitment” process of body proteins has not only been observed in snakes but also in various other venomous animals, such as cone snails, spiders, scorpions and sea anemones as well as hematophagous animals (Fry et al. 2009). Although this overarching concept existed in the field of snake venom toxins for decades, there is not much direct molecular evidence for this process of “recruitment”.

Our laboratory has extensively characterized prothrombin activators from Australian elapid snake venoms and documented their structural and functional similarity with mammalian plasma coagulation factors. Through systematic, detailed studies, we provided the molecular details of the “recruitment” of venom prothrombin activators from plasma coagulation factors after gene duplication. We also identified several key structural changes that make these prothrombin activators better toxins. In this chapter, we will describe the first molecular evidence for the “recruitment” process and the evolution of prothrombin activators in venoms of Australian elapid snakes.

## 2. Snake venom prothrombin activators

As mentioned above, the circulatory system is one of the main targets of snake venom toxins. These toxins affect heart function (e.g., cardiotoxins), vasculature and blood pressure (e.g., sarafotoxins, natriuretic peptides and hemorrhagic toxins), blood coagulation (e.g., procoagulant and anticoagulant proteins), platelet aggregation (e.g., agonists and antagonists) and fibrinolysis (e.g., direct and indirect fibrinolytic proteinases) (Braud et al. 2000; Chow and Kini 2001; Hutton and Warrell 1993; Kini 2004; Kini and Chow 2001; Kini and Evans 1990; Markland 1997; Markland 1998; Morita 2004). All procoagulant proteins are generally proteinases that promote blood coagulation by activating zymogens of specific plasma coagulation factors (Davie 2003). Prothrombin activators are a group of procoagulant proteins that specifically activate prothrombin to thrombin, which then induces blood coagulation through fibrin clot formation.

Several prothrombin activators have been identified and characterized from snake venoms (Gao et al. 2002; Hasson et al. 2003; Joseph and Kini 2001; Kornalik and Blomback 1975; Morita and Iwanaga 1978; Rosing and Tans 1991; Rosing and Tans 1992; Schieck et al. 1972; Silva et al. 2003; Speijer et al. 1986; St Pierre et al. 2005; Yamada and Morita 1997), and based on their properties (structure, cofactor requirements and end-products formed), these snake venom prothrombin activators have been classified into four groups (Kini et al. 2001) (Table 1). Group A and B prothrombin activators, such as ecarin from *Echis carinatus* venom (Kornalik and Blomback 1975; Morita and Iwanaga 1978; Nishida et al. 1995; Schieck et al. 1972) and multactivase from *E. multisquamatus* venom (Yamada and Morita 1997), are metalloproteinases which induce coagulation by converting prothrombin to meizothrombin. These proteins are structurally distinct from blood coagulation factors. Group C and D prothrombin activators, such as oscutarin from *Oxyranus scutellatus* venom (Owen and Jackson 1973; Speijer et al. 1986; Walker et al. 1980; Welton and Burnell 2005) and notecarin from *Notechis scutatus scutatus* venom (Tans et al. 1985), are serine proteinases that induce coagulation by converting prothrombin to thrombin. These proteins exhibit functional similarity to mammalian plasma coagulation factors. However, no detailed structural information of these proteins was available. To fill this void, we initiated structural studies of these prothrombin activators. We characterized one representative each of group C (pseutarin C) and group D (trocarin D) prothrombin activators. Our results show that snake venom prothrombin activators are structurally and functionally similar to mammalian plasma coagulation factors.

Group	Cofactor Requirements	Type of proteinase	Product formed	Estimated Size	Similar to Plasma Coagulation Factors	Examples
A	None	Metalloproteinase	Meizothrombin	~47 kDa	None	Ecarin
B	Ca <sup>2+</sup>			Two subunits of ~25 and ~60 kDa		Carinactivase, multactivase
C	Ca <sup>2+</sup> plus phospholipids	Serine proteinase	Thrombin	Two subunits of ~60 and ~220 kDa	"FXa-FVa" complex	Pseutarin C, oscutarin,
D	Ca <sup>2+</sup> plus phospholipids plus factor Va			~60 kDa	FXa	Trocarin D, hopsarin D, notanarin D

Table 1. Classification of snake venom prothrombin activators

## 2.1 Group D prothrombin activators

Group D prothrombin activators are found exclusively in the venom of Australian elapid snakes (Rosing and Tans 1991). Notecarin from *Notechis scutatus scutatus* venom was the first member of this group to be isolated and characterized (Tans et al. 1985). Since then, similar prothrombin activators have been characterized from several other snake venoms. They are glycoproteins with a molecular weight of ~50 kDa (Table 1). As a group of proteins, they share striking resemblances and requirements for optimal activity with activated mammalian plasma coagulation factor X (FXa) (Table 1) (Joseph et al. 1999; Marsh et al. 1997; Rao and Kini 2002; Stocker et al. 1994; Tans et al. 1985).

Venom of the Australian elapid *Tropidechis carinatus* (rough-scaled snake) was documented to have procoagulant properties 29 years ago (Chester and Crawford 1982). A prothrombin activator was isolated using gel filtration and benzamidine-based affinity chromatography and was partially characterized (Marsh et al. 1997). Our laboratory purified a prothrombin activator, trocarin D, from *T. carinatus* venom to homogeneity using a series of high performance liquid chromatography techniques including gel filtration, ion-exchange and reverse-phase chromatographies (Joseph et al. 1999). This purification procedure was refined to a single-step reverse-phase chromatographic method and was subsequently used for the purification of several other group D prothrombin activators such as notanarin D from *N. ater niger* venom, notecarin D from *N. scutatus* venom and hopsarin D from *Hoplocephalus stephensi* venom (Rao et al. 2003a). Our laboratory characterized trocarin D for its functional and structural properties in detail as a representative of group D prothrombin activators.

Functionally, trocarin D has properties which are similar to mammalian plasma coagulation FXa. They both promote blood coagulation by activating prothrombin to thrombin (Joseph et al. 1999). Trocarin D and FXa achieve activation of prothrombin by cleaving the same peptide bonds (Arg<sub>274</sub>-Thr<sub>275</sub> and Arg<sub>323</sub>-Ile<sub>324</sub>). Both proteins have identical co-factor requirements of Ca<sup>2+</sup> ions, phospholipids and activated factor V (FVa) for their optimal activities (Joseph et al. 1999). We determined the amino acid sequence of trocarin D and its precursor using Edman degradation (Joseph et al. 1999) and cDNA sequencing (Reza et al. 2005a), respectively. Trocarin D and mammalian FXa share significant sequence identity (~53-60%) and exhibit identical domain architecture (Joseph et al. 1999; Rao et al. 2003a) (Figure 1). Both proteins comprise two chains: a heavy chain, which has a serine proteinase with the characteristic catalytic triad (His<sub>42</sub>, Asp<sub>88</sub> and Ser<sub>185</sub>), and a light chain, which has a Gla domain followed by two epidermal-growth factor-like domains (EGF-I and EGF-II). These two chains are held together by a single inter-chain disulfide bond (Joseph et al. 1999) (Figure 1). The differences between trocarin D and mammalian FXa reside in an insertion in the heavy chain, the size of the activation peptide and post-translational modifications. Firstly, there is a 12-residue insert in the heavy chain of trocarin D (Reza et al. 2005a). However, the functional importance of this insertion is not clear. Secondly, the activation peptide of trocarin D precursor is only 27 residues long (Reza et al. 2005a) compared to the activation peptides of mammalian FXs which ranges from 48 to 52 residues (Figure 1). Lastly, post-translational modifications show that trocarin D is glycosylated, but mammalian FXa is not. In addition, trocarin D also contains a O-linked carbohydrate at Ser<sub>52</sub> of the light chain and a N-linked carbohydrate at Asn<sub>45</sub> of the heavy chain (Joseph et al. 1999) (Figure 1). Interestingly, the O-linked carbohydrate moiety has a N-acetylglucosamine moiety, which is found commonly in nuclear and cytoplasmic proteins but rarely in secreted proteins (Hanover et al. 1987; Holt et al. 1987; Holt and Hart 1986; Snow et al. 1987). The

The diagram illustrates the domain organization of three serine proteases: Mammalian FX, Trocarin D precursor, and PCCS precursor. Each protein consists of a light chain and a heavy chain connected by a disulfide bond (s-s).

- Mammalian FX:** The light chain contains a Gla domain (blue) and two EGF repeats (EGF-I and EGF-II, cyan). The heavy chain contains a Serine proteinase domain (red) and an activation peptide (green, 48-52 residues) that is cleaved from the heavy chain. The Gla domain is flanked by YYYYYY and XXXXXX sequences.
- Trocarin D precursor:** The light chain contains a Gla domain (blue) and two EGF repeats (EGF-I and EGF-II, cyan). The heavy chain contains a Serine proteinase domain (red) and an activation peptide (green, 27 residues) that is cleaved from the heavy chain. The Gla domain is flanked by YYYYYY and XXXXXX sequences. A triangle indicates a cleavage site between the Gla domain and EGF-I.
- PCCS precursor:** The light chain contains a Gla domain (blue) and two EGF repeats (EGF-I and EGF-II, cyan). The heavy chain contains a Serine proteinase domain (red) and an activation peptide (green, 27 residues) that is cleaved from the heavy chain. The Gla domain is flanked by YYYYYY and XXXXXX sequences. A triangle indicates a cleavage site between the Gla domain and EGF-I.

The group C snake venom prothrombin activators are also found exclusively in the venoms of Australian elapid snakes (Rosing and Tans 1991). Oscutarin from *Oxyuranus scutellatus* venom was the first member of the group C prothrombin activators to be isolated and characterized (Owen and Jackson 1973; Speijer et al. 1986; Walker et al. 1980; Welton and Burnell 2005). This group of proteins is generally ~300 kDa in size and comprises two subunits (~60 and ~220 kDa) (Table 1). The smaller enzymatic subunit is a serine proteinase and has characteristics of FXa, whereas the nonenzymatic subunit resembles activated mammalian plasma coagulation factor V (FVa). Overall, group C prothrombin activators have striking resemblances to and similar co-factor requirements as the mammalian plasma coagulant “FXa-FVa” complex (Filippovich et al. 2005; Masci et al. 1988; Rao and Kini 2002; Speijer et al. 1986; Walker et al. 1980) (Table 1).

Pseutarin C was purified from *P. textilis* venom, and it activates prothrombin to thrombin. For its optimal activity, pseutarin C requires only  $\text{Ca}^{2+}$  ions and phospholipids (Rao and Kini 2002). These functional characteristics are similar to that of the mammalian "FXa-FVa" complex. As with other group C prothrombin activators, pseutarin C comprises two subunits of ~60 kDa and ~220 kDa (Rao and Kini 2002) (Table 1). The smaller subunit, with serine proteinase activity, was termed the pseutarin C catalytic subunit (PCCS) and the larger subunit, which has no enzymatic activity, was termed the pseutarin C nonenzymatic subunit (PCNS) (Rao and Kini 2002). A comparison of the protein quantities in the venom and plasma revealed that pseutarin C is expressed ~4,200 times higher in the venom than the amount of FV and FX in the plasma (Rao and Kini 2002). We purified pseutarin C and its subunits and characterized them both functionally and structurally as representatives of the group C prothrombin activators.

1. *Pseutarin C catalytic subunit (PCCS)* – Functionally, PCCS is similar to mammalian FXa and group D prothrombin activators (Rao and Kini 2002). They have the same co-factor requirements, including  $\text{Ca}^{2+}$  ions, phospholipids and FVa, for their optimal activity and activate prothrombin by cleaving the same two peptide bonds (Arg<sub>274</sub>-Thr<sub>275</sub> and Arg<sub>323</sub>-Ile<sub>324</sub>). The enzymatic activity ( $V_{\text{max}}$ ) of PCCS is enhanced by the presence of FVa (Rao and Kini 2002). The amino acid sequence of PCCS and its precursor was determined using both Edman degradation and cDNA sequencing (Rao et al. 2004; Rao and Kini 2002). Structurally, PCCS is also similar to mammalian FXa and group D prothrombin activators (Figure 1). Its sequence shows ~42% identity to mammalian FXa and 74-83% identity to group D prothrombin activators (Rao et al. 2004). Like mammalian FXa and group D prothrombin activators (Rao et al. 2004), the domain architecture of PCCS consists of a light and a heavy chain that are linked by a single disulfide bond (Rao et al. 2004). The light chain has a Gla domain followed by two EGF-like domains, and the heavy chain contains a serine proteinase domain (Figure 1). Despite such functional and structural similarities, the differences between PCCS and mammalian FXa reside in the size of the activation peptide and post-translational modifications (Rao et al. 2004). Similar to trocarin D precursor, the activation peptide of PCCS precursor is 27 residues long and is significantly shorter than those of mammalian FXs (Figure 1). Like trocarin D, it has an insertion in its heavy chain. Interestingly, the PCCS insert is 13 residues long and is distinctly different from the 12-residue insert in trocarin D (Rao et al. 2004; Reza et al. 2005b). This strongly indicates that the evolution of groups C and D prothrombin activators are independent. Like trocarin D, the Ser<sub>52</sub> and Asn<sub>45</sub> residues of the light and heavy chains of PCCS are O- and N-glycosylated, respectively (Figure 1). However, as mentioned previously, these two residues have no post-translational modifications in mammalian FX/FXa. These functional and structural characteristics suggest that PCCS is a homologue of mammalian FXa and group D prothrombin activators.

2. *Pseutarin C nonenzymatic subunit (PCNS)* – Structurally, PCNS is similar to mammalian FV. The amino acid sequence of PCNS and its precursor was determined using Edman degradation and cDNA sequencing (Rao et al. 2003b; Rao and Kini 2002). PCNS shares ~50% identity with the mammalian FV and has identical domain architecture with mammalian FV (Rao et al. 2003b). Both PCNS and mammalian FV have six domains: A1, A2, B, A3, C1 and C2 (Figure 2). Domains A and C are functionally important, and these domains are highly conserved in PCNS and FVs of other species (Rao et al. 2003b). These structural similarities suggest that PCNS is a homologue of mammalian FV (Rao et al. 2003b).

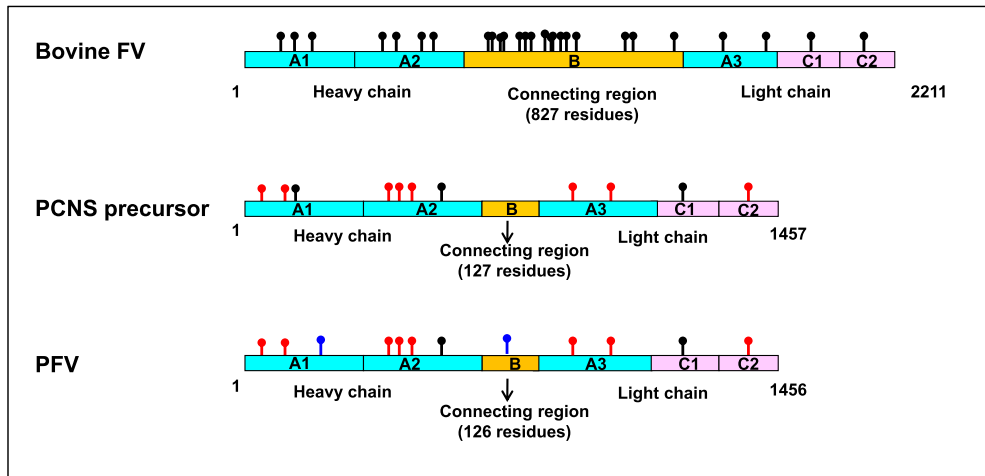


Fig. 2. Domain architecture of bovine FV, PCNS precursor and PFV. N-glycosylation sites are shown as colored knobs. Black knobs are conserved, while red knobs are found only in PCNS and PFV, and blue knobs are found only in PFV.

Despite being a FV homologue, PCNS shows several differences with FV from other species. Firstly, the domain B size of PCNS is significantly smaller (127 residues) than that of fishes (fugu: 530 residues and zebrafish: 756 residues) and mammals (murine: 843 residues, bovine: 869 residues, and human: 882 residues) (Rao et al. 2003b) (Figure 2). During FV activation, domain B is removed by thrombin or FXa by cleavage at three activation sites: Arg<sub>709</sub>, Arg<sub>1018</sub> and Arg<sub>1545</sub> (bovine FV numbering) (Foster et al. 1983; Nesheim et al. 1979; Suzuki et al. 1982) (Figure 3B). Although only two of these sites (Arg<sub>709</sub> and Arg<sub>1545</sub>) are conserved in PCNS (Figure 3B), the complete domain B can still be cleaved off during the activation of PCNS (Rao et al. 2003b). Thus, the difference in domain B size should not have any effect on the function of PCNS. Secondly, PCNS and FV of other species have different post-translation modifications. While bovine FV has 29 glycosylation sites, PCNS has only 11 potential N-glycosylation sites (Rao et al. 2003b) (Figure 2). Mammalian FV is phosphorylated at the Ser<sub>692</sub> residue, but PCNS is not (Rao et al. 2003b). In addition, there are six sulfation sites in human FV that are absent in PCNS (Rao et al. 2003b). This difference in post-translation modifications is interesting, as sulfation and phosphorylation are important for regulating the activation of human FV by thrombin (Kalafatis et al. 1994; Pittman et al. 1994).

Aside from these differences, it is noted that PCNS possesses certain modifications and associations that allow it to function efficiently as a toxin (Bos et al. 2009). Firstly, PCNS has evolved a way to evade inactivation by protein C. In the human coagulation system, protein C is activated by thrombin in the presence of thrombomodulin, and activated protein C (APC) subsequently inactivates FVa in a negative feedback loop (Esmon 2001). This inactivation occurs by proteolytic cleavage at three sites on the FVa heavy chain: Arg<sub>306</sub>, Arg<sub>506</sub> and Arg<sub>662</sub> (Kalafatis et al. 1994; Mann et al. 1997) in bovine FVa (Figure 3B). PCNS is able to evade APC inactivation, as it does not have any of the three APC cleavage sites (Rao et al. 2003b). Even an alternate less efficient cleavage site at Arg<sub>316</sub> (van der Neut et al. 2004b)

is not conserved in PCNS (Rao et al. 2003b). Secondly, FVa is inactivated by phosphorylation at Ser<sub>692</sub> (Kalafatis et al. 1994). This phosphorylation site is not present in PCNS (Rao et al. 2003b). Lastly, PCNS is shielded from APC inactivation (Nesheim et al. 1982; Rao et al. 2003b) and is kept activated through its constant and stable association with FXa-like PCCS (Rao et al. 2003a; Rao et al. 2004; Thorelli et al. 1998) in the venom. We have shown experimentally that pseutarin C is unaffected by APC, while bovine FXa-FVa complex is completely inactivated (Bos et al. 2009; Rao et al. 2003b) (Figure 4). Overall, PCNS is a good example of how a toxin gene is duplicated from an ancestral gene and undergoes modifications to gain unique characteristics that allow it to function efficiently as a toxin.

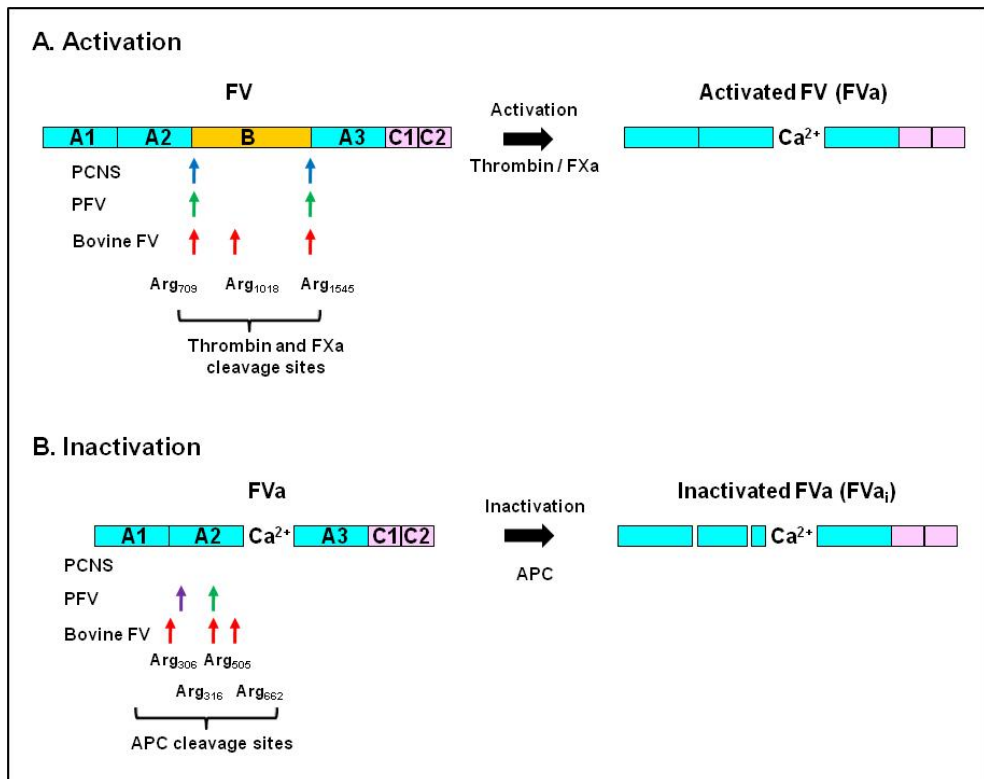


Fig. 3. Functionally important proteolytic sites in PCNS and FV. (A) Activation sites. Thrombin and FXa cleavage sites of PCNS, PFV and bovine FV are shown in blue, green and red, respectively. Critical activation sites are conserved in PCNS and PFV. (B) Inactivation sites. Activated protein C (APC) cleavage sites of PFV and bovine FV are shown in green and red, respectively. In addition to Arg<sub>505</sub>, PFV has a primitive inactive site at Arg<sub>316</sub> shown in purple arrow. Inactivation sites are missing in PCNS.

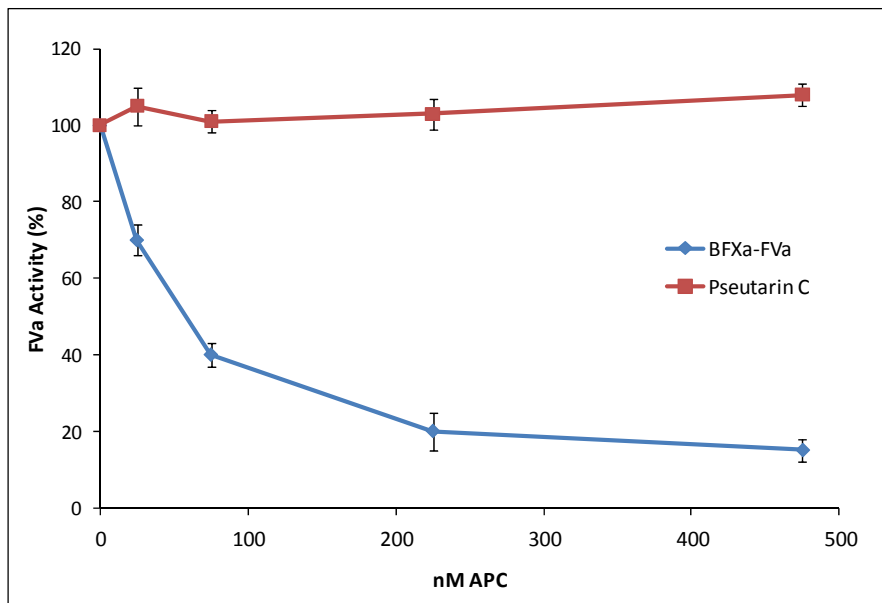


Fig. 4. APC resistance assay (Rao et al. 2003b). Varying concentrations of APC were added either to pseutarin C (E; 8 nM) or bovine FXa-FVa (F; FXa 42 nM, FVa 2 nM) complex which was diluted in 50 mM Tris-HCl buffer (pH 7.5) containing 100 mM NaCl, 5 mM CaCl<sub>2</sub>, and 0.5 mg/mL BSA. The reaction mixture was incubated for 30 minutes at room temperature. Prothrombin was added to a final concentration of 2.8  $\mu$ M and thrombin formed was assayed using thrombin-specific chromogenic substrate S-2238. Each point represents an average of 2 independent experiments each carried out in triplicates.

### 3. Parallel prothrombin activator system in Australian elapid snakes

As described above, groups C and D snake venom prothrombin activators are functional and structural homologues of mammalian blood coagulation factors. As snakes are vertebrates, their hemostatic system should contain plasma coagulation factors. Thus, Australian elapid snakes should possess parallel prothrombin activating systems: one in their venom, which is used as an offensive weapon to attack the hemostatic system of the prey, and the other in their plasma, which is used for their own hemostatic purpose. We examined the presence of plasma coagulation factors in the snake's hemostatic system and determined the relationship between the snake venom and plasma coagulation factors.

#### 3.1 Trocarin D and FX from *Tropidechis carinatus* (TrFX)

Since the liver mainly expresses plasma coagulation factors, the cDNA encoding *T. carinatus* FX (TrFX) was sequenced from liver tissue (Reza et al. 2005a). The deduced amino acid sequence of TrFX is similar to mammalian FX (~50%) and trocarin D (~80%) (Figure 5). Structurally, TrFX is similar to trocarin D. They both have conserved cysteine residues and identical domain architecture. However, there are some differences between TrFX and trocarin D. The activation peptide of TrFX is similar to the mammalian FXs and not to that of



the venom prothrombin activators (Reza et al. 2005b). It is 57 residues long compared to 27 residues in trocarin D (Figure 5). In addition, there is no 12-residue insert in the heavy chain of TrFX as was observed to be present in the trocarin D precursor (Reza et al. 2005b). These differences in amino acid sequences, and the lengths of activation peptides and insertion in the heavy chain, suggest that TrFX and trocarin D are encoded by two independent genes. Hence, this confirms the presence of a parallel prothrombin activator system. TrFX is more similar to trocarin than to mammalian FX in terms of post-translational modifications (Reza et al. 2005a). TrFX and trocarin D both have *N*- and *O*-glycosylation modifications that are not found in mammalian FXs (as described previously).

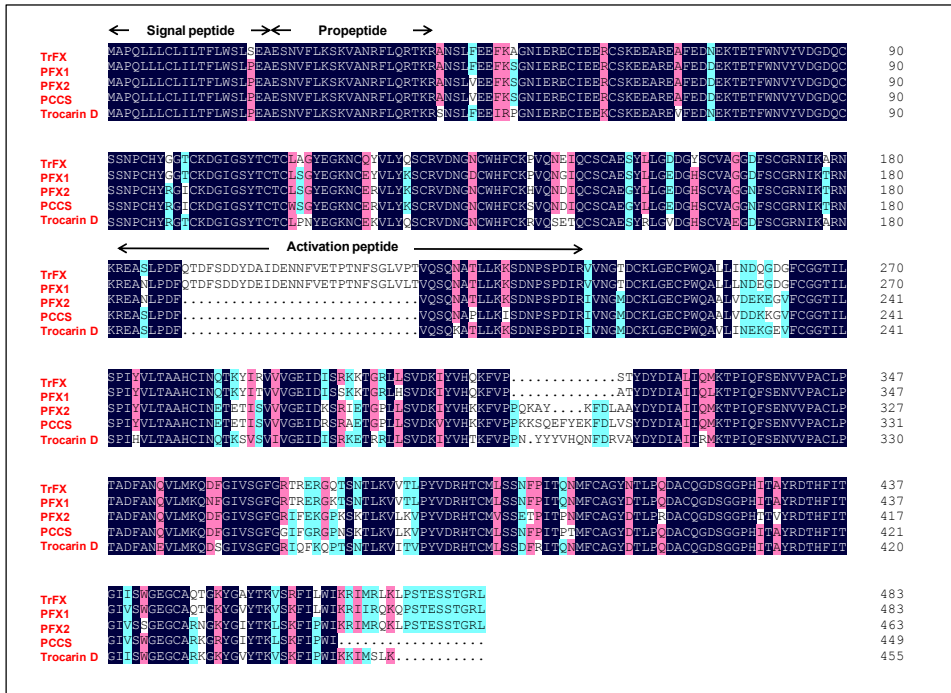


Fig. 5. Alignment of deduced amino acid sequences of FX-like proteins from *T. carinatus* (TrFX and trocarin D) and *P. textilis* (PCCS, PFX1 and PFX2) snakes.

Trocarin D and TrFX differ in their physiological roles. Trocarin D plays an offensive role as a toxin in the venom that is used for killing prey. Upon envenomation, like other prothrombin activators (Masci et al. 1988; Rao et al. 2003a), it induces cyanation and death in experimental animals (Joseph et al. 1999) through disseminated intravascular coagulopathy. On the other hand, TrFX plays a crucial role in the coagulation cascade and prevents excessive blood loss by promoting blood coagulation when there is a vascular injury. Trocarin D is an active enzyme and is found in large quantities in the venom. In contrast, TrFX is found as a zymogen, which gets activated only when required and is found in much smaller concentrations in the plasma. Real-time polymerase chain reaction (RT-PCR) was used to determine the amount of expression of these two closely related proteins in the liver and venom gland. The results indicate that trocarin D is expressed in the venom gland but

not in the liver, while TrFX is expressed in the liver but not in the venom gland. Further, the expression of trocarin D is ~30 times higher in the venom gland than TrFX in the liver (Reza et al. 2007). Such differential expression patterns of trocarin D and TrFX strongly support the distinct physiological roles of these two proteins.

### 3.2 PCCS and FX from *Pseudonaja textilis* (PFX)

To understand the evolution of group C prothrombin activators, we also determined the cDNA sequence of the *P. textilis* FX (PFX) from the liver. Interestingly, two PFX isoforms (PFX1 and PFX2) were detected in the liver, and their cDNA sequences are ~85% similar (Reza et al. 2006). The domain architecture and cysteine residues of these two isoforms are also conserved compared to group D prothrombin activators. Amino acid sequence comparison shows that PFX1 is more similar to TrFX (~94%), while PFX2 is more similar PCCS and trocarin D (~90%) (Figure 5). Further, PFX1 has a longer activation peptide, similar to plasma FXs, whereas PFX2 has a shorter activation peptide, similar to PCCS and trocarin D. Also, PFX2 has a 9-residue insert, which is not present in PFX1. These structural differences suggest that PFX1, PFX2 and PCCS are encoded by three independent genes and that PFX2 is an evolutionary intermediate between PFX1 and PCCS (Reza et al. 2006) (Figure 5). This similarly confirms the presence of a parallel prothrombin activator system. The expression profiles of PFX1, PFX2 and PCCS were determined in liver and venom gland tissues by RT-PCR (Reza et al. 2006). The results show that PFX1 and PFX2 are expressed only in the liver, while PCCS is expressed only in the venom gland. PFX1 is also found to be expressed ~55,000 times higher than PFX2 in the liver, and PCCS is expressed ~80 times higher in the venom gland than is PFX1 in the liver (Reza et al. 2006). In summary, the sequence comparisons and expression profiles indicate that PCCS has evolved from PFX1 by gene duplication and PFX2 is an intermediary product of this "recruitment" process (Figure 6).

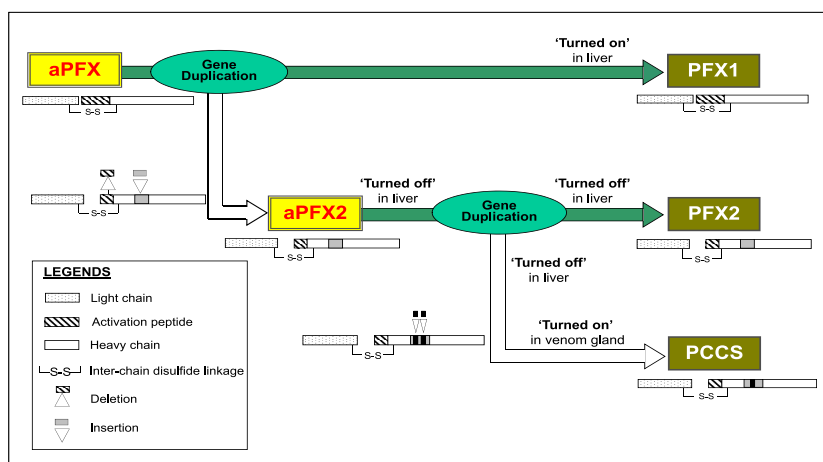


Fig. 6. Schematic diagram showing the probable evolutionary path in the recruitment of FX protein as toxin in the venom (Reza et al. 2006).

### 3.3 PCNS and FV from *Pseudonaja textilis* (PFV)

The cDNA sequence of *P. textilis* FV (PFV) was determined from its liver (Minh et al. 2005). The deduced amino acid sequence of PFV shows similarities to other mammalian and non-mammalian FVs (~50%) and PCNS (~96%) and shares identical domain architecture (Minh et al. 2005). Like the FVs of other species, PFV and PCNS comprise A1, A2, B, A3, C1 and C2 domains (Figure 3A). Functionally important domains A and C are highly conserved in both PFV and PCNS, whereas domain B is the most variable (Minh et al. 2005). The domain B (126 residues) of PFV is one residue shorter than that of PCNS (127 residues), and is much shorter than that of mammalian and non-mammalian FVs. A more detailed comparison shows that all the FXa and thrombin proteolytic cleavage sites (which are important for activation of these nonenzymatic proteins) are conserved in PFV and PCNS (Figure 3A). However, PFV has an additional FXa proteolytic cleavage site at Arg<sub>1765</sub> (Minh et al. 2005; Rao et al. 2003b). This cleavage site also exists in mammalian FV but not in FVs of teleosts (Minh et al. 2005). This is evolutionarily interesting as this additional cleavage site may be a characteristic found only in tetrapod FVs. However, the functional implication of this cleavage with regards to procoagulant activity of FV is not yet known.

As mentioned previously, PCNS has evolved to be resistant to inactivation by activated protein C (APC), which is crucial to its function as a toxin. On the other hand, PFV is similar to other FV, as it can still be inactivated by APC. PFV can be inactivated by APC by cleavage at Arg<sub>316</sub>, a primitive inactive site (van der Neut et al. 2004a), and at Arg<sub>506</sub> (Minh et al. 2005) (Figure 3B). The expression profiles of PFV and PCNS in the liver and the venom gland were determined using RT-PCR. As with other venom prothrombin activator genes, PCNS is expressed only in the venom gland, while PFV is expressed only in the liver. It was found that PCNS is expressed ~280 times higher in the venom gland than is PFV in the liver (Minh et al. 2005). Thus, PCNS and PFV have differential expressions (Minh et al. 2005).

Based on sequence comparisons, we confirmed the presence of parallel prothrombin activator systems in Australian elapid snakes and showed for the first time that groups C and D prothrombin activators in snake venom and their plasma coagulation factor counterparts are closely related. We also proposed that these venom prothrombin activators evolved from their plasma coagulation factor counterparts by gene duplication and were subsequently modified to function efficiently as toxins.

## 4. Phylogenetic relationship between snake venom and plasma prothrombin activators

A phylogenetic tree of the snake venom and plasma prothrombin activators with other known FX sequences was constructed to understand their evolutionary relationships using zebrafish FX as the out group (Reza et al. 2006; St Pierre et al. 2005). All the reptilian sequences form a monophyletic group (Reza et al. 2006) (Figure 7). Within the reptilian clade, group C and D prothrombin activators appear as two separate clades on the tree. This indicates that, despite their similarities, group C and D prothrombin activators have originated independently. Interestingly, the PFX2 sequence is found nested within the group C prothrombin activators. This supports the hypothesis that PFX2 is an evolutionary intermediate of PCCS from PFX1. Based on the topology of the phylogenetic tree, it is suggested that these snake venom prothrombin activators have been “recruited” through independent evolutionary events (Reza et al. 2006) (Figure 7).

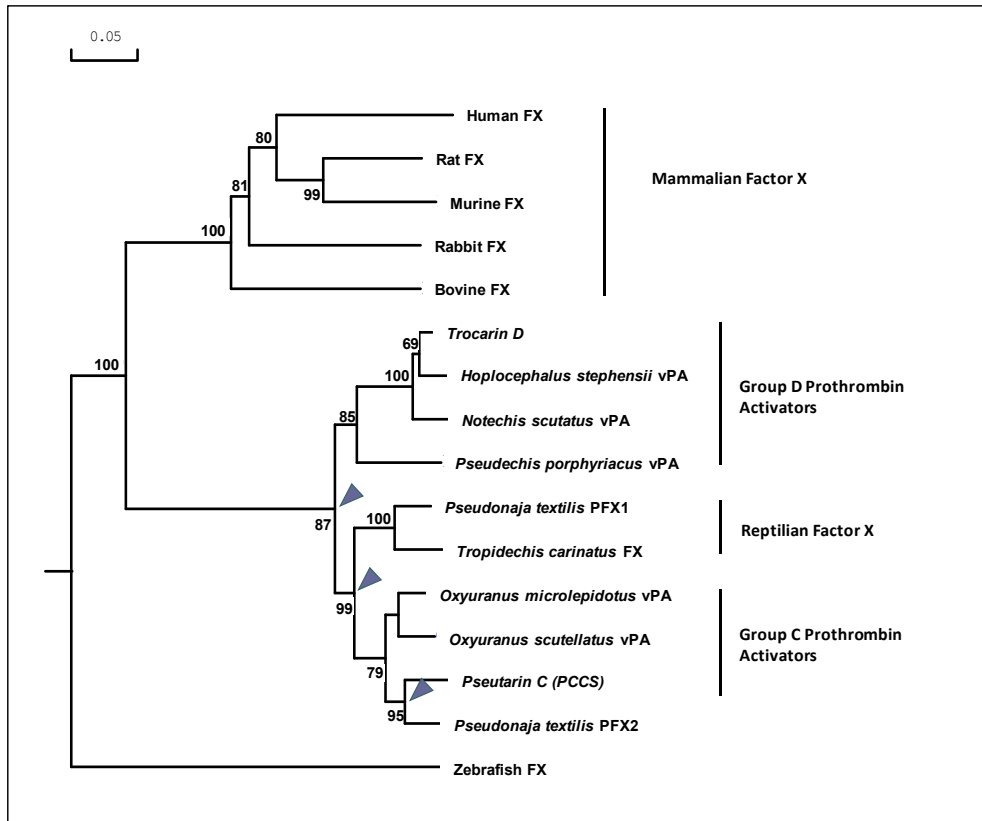


Fig. 7. Phylogenetic relationships of snake venom and plasma prothrombin activators with other known FX sequences (Reza et al. 2006). “vPA” is an abbreviation for venom prothrombin activators. Arrows indicate the three independent “recruitment” events of snake venom prothrombin activators.

## 5. Comparison of trocarin D and TrFX genes

In the previous sections, we have described how the venom prothrombin activators have been modified to gain certain characteristics, such as resistance to inactivation, which enables them to function better as toxins. However, differential and tissue-specific expression of venom prothrombin activators and their plasma coagulation factors is also important for their respective physiological roles. The expression of toxins should be venom gland-specific and inducible to higher levels. This is so the snake can protect itself against its own venom toxins and replenish its venom supply quickly. Conversely, plasma coagulant factors are mainly expressed in the liver at constitutently low levels so that they can be activated to induce blood coagulation during vascular injuries.

To understand how the venom prothrombin activators are regulated for tissue specificity and level of expression, we determined the gene structure of trocarin D and TrFX. Based on the cDNA sequences of trocarin D and TrFX, and that of mammalian FX gene, primers were

designed, and the gene sequences were determined using genomic DNA PCR and genome walking strategies (Reza et al. 2006). The gene organizations of trocarin D and TrFX are identical (eight exons and seven introns). The intron sequences were highly similar to each other with only differences in the promoter and intron 1 regions (Figure 8). Such similarities strongly support our findings that there are two closely related, parallel prothrombin activator systems, and that the venom prothrombin activators are “recruited” from plasma coagulation factors through gene duplication. The duplicated FX gene was subsequently modified and “recruited” for expression in the venom gland as a venom prothrombin activator.

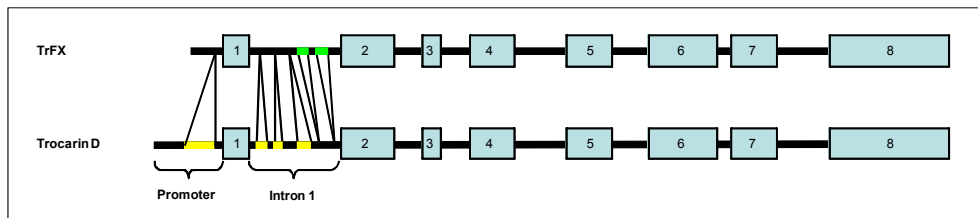


Fig. 8. Gene organization comparison of trocarin D and TrFX. Exons are shown as numbered boxes. The differences in the promoter and intron 1 regions are indicated in yellow (insertions in trocarin D gene) and green (deletions in TrFX gene) (Reza et al., 2007).

### 5.1 *Cis*-elements in trocarin D promoter region

The overlapping promoter regions of trocarin D and TrFX were characterized by comparing them against previously characterized human (Hung et al. 2001; Hung and High 1996) and murine (Wilberding and Castellino 2000) FX promoter regions (Reza et al. 2007) (Figure 9). Based on these comparisons, four conserved *cis*-regulatory elements in the trocarin D and TrFX promoter regions were identified (Figure 9): (i) a CCAAT box (Hung et al. 2001; Hung and High 1996; Wilberding and Castellino 2000), (ii) a gut-specific transcription factor GATA-4 binding site (Hung et al. 2001), (iii) a liver-specific transcription factor HNF-4 (Hung and High 1996), and (iv) multiple Sp1/Sp3 binding sites (Hung et al. 2001).

Comparison of the trocarin D and TrFX promoter regions reveals that trocarin D has a 264 bp insertion (Figure 8 and 9). This 264 bp is located from -33 to -297 bp upstream of the trocarin D start codon (ATG) (Figure 9). This insertion is postulated to play a major role in the recruitment of the duplicated TrFX gene by causing it to be exclusively expressed in the venom gland as the procoagulant toxin, trocarin D. Hence, it was termed Venom Recruitment/Switch Element (*VERSE*). This segment was characterized for its *cis*-elements and gene-regulatory role using luciferase assays in primary venom gland cells and mammalian cell lines (Kwong et al. 2009). The *VERSE* promoter was found to be responsible for the elevation of expression levels, but not tissue-specific expression, of trocarin D. In terms of *cis*-element characterization, besides confirming the presence of two TATA-boxes, one GATA box and one Y-box, three novel *cis*-elements were also identified (Figure 9). Functionally, it is found that both TATA boxes (TLB2 and TLB3) are functional. However, TLB2 is the primary TATA box which initiates and directs transcription start site (Kwong et al. 2009) (Figure 9).

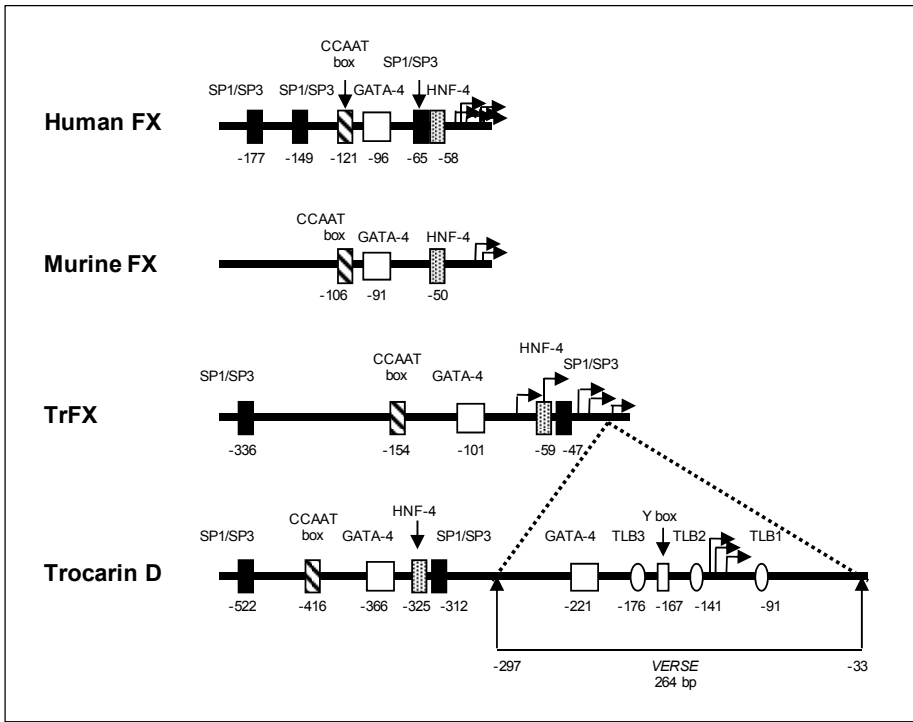


Fig. 9. Comparison of promoter regions in mammalian and *T. carinatus* prothrombin activator genes (Reza et al., 2007).

## 5.2 Comparison of trocarin D and TrFX first introns

The intron 1 size of trocarin D is 7911 bp, while that of TrFX is 5293 bp. The difference in size is explained by three insertions and two deletions in the trocarin D intron 1 region (Reza et al. 2007) (Figure 10). Bioinformatics analysis of these insertion/deletion segments reveals that they are novel. The three insertion segments within intron 1 of trocarin D region are 214 bp, 1975 bp and 2174 bp in size with respective positions at 128 bp, 914 bp and 3300 bp on the trocarin D gene (Figure 10). Upon closer analysis of the insertion segment sequences, it is observed that the first insert within intron 1 of trocarin D is almost an exact repeat (96.33% identity) of the intron 1 segment spanning from 3082 bp to 3299 bp. The other two inserts seem to be inverted repeats of each other (~71% identity). The third insert shows potential of being a Scaffold/Matrix Attachment Region (S/MAR) due to: (i) a high AT content (Cockerill and Garrard 1986; Liebich et al. 2002; Zhou and Liu 2001), (ii) a topoisomerase II (Boulikas 1993), (iii) a S/MAR consensus motif (van Drunen et al. 1997), (iv) a significant over-representation of characteristic hexanucleotides (Liebich et al. 2002), and (v) an ATTA motif and an AT-rich region with H-box (Will et al. 1998). The two deletion segments within intron 1 of trocarin D are 255 bp and 1406 bp in size with respective positions at 2610 bp and 3770 bp on the TrFX gene (Reza et al. 2007) (Figure 10). As the *VERSE* promoter of trocarin D does not regulate tissue-specific expression (Kwong et al. 2009), it is postulated that these insertions and deletions in the

intron 1 of trocarin D could contain *cis*-elements that are responsible for the venom gland-specific expression of trocarin D. In summary, the characterization of the *VERSE* promoter and intron 1 regions of trocarin D has increased our understanding regarding gene regulation of venom prothrombin activators.

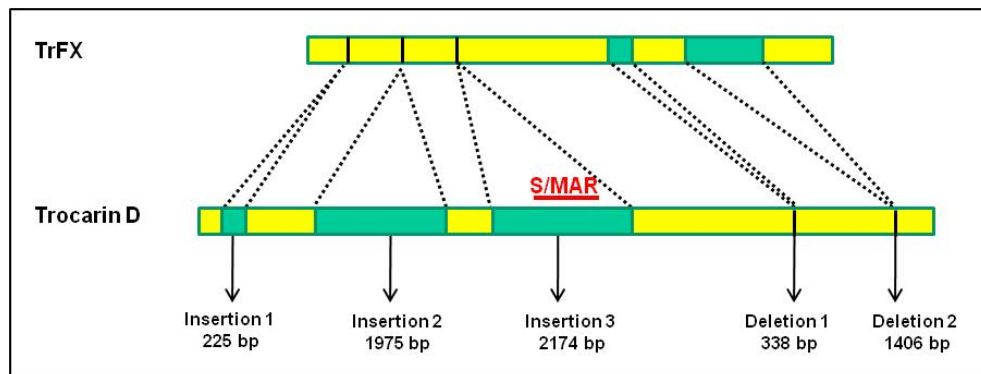


Fig. 10. Comparison of intron 1 regions in TrFX and trocarin D genes (Reza et al., 2007).

## 6. Gene duplication in snake venom toxin diversification

Besides “recruitment”, gene duplication also plays an important role in the diversification of venom toxins. This diversification is essential for the development of novel toxins. This diversification through gene duplication is evident from the many toxin isoforms present in the snake venom. Interestingly, each isoform varies in its function and gene regulation.

Gene duplication has also led to neofunctionalization of venom toxins, which has led to the new families of snake venom toxins (St Pierre et al. 2008) and addition of new members within these families (Fry et al. 2003; Landan et al. 1991b; Lynch 2007; Moura-da-Silva et al. 1996). The three-finger toxin (3FTx) multigene family is a good example of neofunctionalization by gene duplication (Fry et al. 2003). Structurally, all the members of this family have very well-conserved cysteine residues and share a common structure of three beta-stranded loops extending from a central core. However, they exhibit a wide variety of pharmacological effects. For example, acetylcholinesterase inhibition (fasciculin from *Dendroaspis angusticeps* venom), neurotoxicity ( $\alpha$ -bungarotoxin from *Bungarus multicinctus* venom), cardiotoxicity ( $\beta$  cardiotoxin from *Ophiophagus hannah* venom), and many others (for details, see (Kini and Doley 2010)). Neofunctionalization occurs when a toxin gene undergoes gene duplication and the duplicated gene is mutated within the functional sites, which often results in new ligand-binding specificities (Kini 2002).

Besides neofunctionalization, changes in gene regulation are also the outcomes of gene duplication. This can be seen in two isoforms present in the venom of *Naja sputatrix*: cardiotoxin and  $\alpha$ -neurotoxin (Ma et al. 2001). Besides varying in function, these two isoforms have different expression levels in the venom gland. Cardiotoxin constitutes 60% of the venom while the  $\alpha$ -neurotoxin makes up only 3% of the venom. Gene duplication is evident from the gene comparison whereby the structures and amino acid sequence of these

two toxins are very well-conserved (Ma et al. 2002). The main difference lies in the promoter segment, where it was found that the  $\alpha$ -neurotoxin promoter contains a stronger silencer element, which is responsible for significantly reducing its expression level in the venom (Ma et al. 2001; Ma et al. 2002).

In the case of venom prothrombin activators, we have shown that they have been "recruited" from the gene of an ancestral plasma prothrombin activator protein through gene duplication. The duplicated gene underwent modifications in its regulatory and coding regions to gain toxin characteristics. *VERSE* segments were inserted in the promoter regions of trocarin D and PCCS and are responsible for their elevated level of expression. Insertion/deletion segments in their intron 1 regions are postulated to be responsible for venom-gland specific expression. Modifications in the gene-coding regions enable prothrombin activators to function better as toxin by gaining certain characteristics such as resistance to inactivation.

## 7. Conclusion

Gene duplication has played a major role in the development of snake venom toxins. Our findings on venom prothrombin activators and blood coagulation factors have captured the first molecular evidence of gene duplication. The characterization of the differences in their genes, i.e. *VERSE* segment and intron 1 of trocarin D, has increased our understanding of gene regulation of snake venom toxins. It is shown that the *VERSE* segment is responsible for the elevation of gene expression and that the intron 1 is probably responsible venom gland-specific expression (unpublished observations). We identified three novel *cis*-elements in the *VERSE* segment, and these play important roles in gene regulation. It would be interesting to further characterize them and their *trans*-factor partners to determine how various *trans*-factors interact with each other to regulate gene expression. The answers to some of these questions will increase our overall understanding of gene regulation.

## 8. References

- Bos, M.H., M.Boltz, L.St Pierre, P.P.Masci, J.de Jersey, M.F.Lavin, and R.M.Camire. 2009. "Venom factor V from the common brown snake escapes hemostatic regulation through procoagulant adaptations." *Blood*. 114:686-692.
- Boulikas, T. 1993. "Nature of DNA sequences at the attachment regions of genes to the nuclear matrix." *J.Cell Biochem*. 52:14-22.
- Braud, S., C.Bon, and A.Wisner. 2000. "Snake venom proteins acting on hemostasis." *Biochimie*. 82:851-859.
- Chester, A. and G.P.Crawford. 1982. "In vitro coagulant properties of venoms from Australian snakes." *Toxicon*. 20:501-504.
- Chow, G. and R.M.Kini. 2001. "Exogenous factors from animal sources that induce platelet aggregation." *Thromb.Haemost*. 85:177-178.
- Cockerill, P.N. and W.T.Garrard. 1986. "Chromosomal loop anchorage of the kappa immunoglobulin gene occurs next to the enhancer in a region containing topoisomerase II sites." *Cell*. 44:273-282.



- Davie, E.W. 2003. "A brief historical review of the waterfall/cascade of blood coagulation." *J.Biol.Chem.* 278:50819-50832.
- Esmon, C.T. 2001. "Role of coagulation inhibitors in inflammation." *Thromb.Haemost.* 86:51-56.
- Filippovich, I., N.Sorokina, L.St Pierre, S.Flight, J.de Jersey, N.Perry, P.P.Masci, and M.F.Lavin. 2005. "Cloning and functional expression of venom prothrombin activator protease from *Pseudonaja textilis* with whole blood procoagulant activity." *Br.J.Haematol.* 131:237-246.
- Foster, W.B., M.E.Nesheim, and K.G.Mann. 1983. "The factor Xa-catalyzed activation of factor V." *J.Biol.Chem.* 258:13970-13977.
- Fry, B.G. 2005. "From genome to "venome": molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins." *Genome Res.* 15:403-420.
- Fry, B.G., K.Roelants, D.E.Champagne, H.Scheib, J.D.Tyndall, G.F.King, T.J.Nevalainen, J.A.Norman, R.J.Lewis, R.S.Norton, C.Renjifo, and R.C.de la Vega. 2009. "The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms." *Annu.Rev.Genomics Hum.Genet.* 10:483-511.
- Fry, B.G., W.Wuster, R.M.Kini, V.Brusic, A.Khan, D.Venkataraman, and A.P.Rooney. 2003. "Molecular evolution and phylogeny of elapid snake venom three-finger toxins." *J.Mol.Evol.* 57:110-129.
- Gao, R., R.M.Kini, and P.Gopalakrishnakone. 2002. "A novel prothrombin activator from the venom of *Micropechis ikaheka*: isolation and characterization." *Arch. Biochem. Biophys.* 408:87-92.
- Hanover, J.A., C.K.Cohen, M.C.Willingham, and M.K.Park. 1987. "O-linked N-acetylglucosamine is attached to proteins of the nuclear pore. Evidence for cytoplasmic and nucleoplasmic glycoproteins." *J.Biol.Chem.* 262:9887-9894.
- Hasson, S.S., R.D.Theakston, and R.A.Harrison. 2003. "Cloning of a prothrombin activator-like metalloproteinase from the West African saw-scaled viper, *Echis ocellatus*." *Toxicon.* 42:629-634.
- Holt, G.D., R.S.Haltiwanger, C.R.Torres, and G.W.Hart. 1987. "Erythrocytes contain cytoplasmic glycoproteins. O-linked GlcNAc on Band 4.1." *J.Biol.Chem.* 262:14847-14850.
- Holt, G.D. and G.W.Hart. 1986. "The subcellular distribution of terminal N-acetylglucosamine moieties. Localization of a novel protein-saccharide linkage, O-linked GlcNAc." *J.Biol.Chem.* 261:8049-8057.
- Hung, H.L. and K.A.High. 1996. "Liver-enriched transcription factor HNF-4 and ubiquitous factor NF-Y are critical for expression of blood coagulation factor X." *J.Biol.Chem.* 271:2323-2331.
- Hung, H.L., E.S.Pollak, R.D.Kudaravalli, V.Arruda, K.Chu, and K.A.High. 2001. "Regulation of human coagulation factor X gene expression by GATA-4 and the Sp family of transcription factors." *Blood.* 97:946-951.
- Hutton, R.A. and D.A.Warrell. 1993. "Action of snake venom components on the haemostatic system." *Blood Rev.* 7:176-189.

- Inoue, K. and T.Morita. 1993. "Identification of O-linked oligosaccharide chains in the activation peptides of blood coagulation factor X. The role of the carbohydrate moieties in the activation of factor X." *Eur.J.Biochem.* 218:153-163.
- Joseph, J.S., M.C.Chung, K.Jeyaseelan, and R.M.Kini. 1999. "Amino acid sequence of trocarin, a prothrombin activator from *Tropidechis carinatus* venom: its structural similarity to coagulation factor Xa." *Blood.* 94:621-631.
- Joseph, J.S. and R.M.Kini. 2001. "Snake venom prothrombin activators homologous to blood coagulation factor Xa." *Haemostasis.* 31:234-240.
- Kalafatis, M., M.D.Rand, and K.G.Mann. 1994. "The mechanism of inactivation of human factor V and human factor Va by activated protein C." *J.Biol.Chem.* 269:31869-31880.
- Kini, R.M. 2002. "Molecular moulds with multiple missions: functional sites in three-finger toxins." *Clin.Exp.Pharmacol.Physiol.* 29:815-822.
- Kini, R.M. 2004. "Platelet aggregation and exogenous factors from animal sources." *Curr.Drug Targets.Cardiovasc.Haematol.Disord.* 4:301-325.
- Kini, R.M. and G.Chow. 2001. "Exogenous inhibitors of platelet aggregation from animal sources." *Thromb.Haemost.* 85:179-181.
- Kini, R.M. and R.Doley. 2010. "Structure, function and evolution of three-finger toxins: mini proteins with multiple targets." *Toxicon.* 56:855-867.
- Kini, R.M. and H.J.Evans. 1990. "Effects of snake venom proteins on blood platelets." *Toxicon.* 28:1387-1422.
- Kini, R.M., T.Morita, and J.Rosing. 2001. "Classification and nomenclature of prothrombin activators isolated from snake venoms." *Thromb.Haemost.* 86:710-711.
- Kornalik, F. and B.Blomback. 1975. "Prothrombin activation induced by Ecarin - a prothrombin converting enzyme from *Echis carinatus* venom." *Thromb.Res.* 6:57-63.
- Kwong, S., A.E.Woods, P.J.Mirtschin, R.Ge, and R.M.Kini. 2009. "The recruitment of blood coagulation factor X into snake venom gland as a toxin: the role of promoter Cis-elements in its expression." *Thromb.Haemost.* 102:469-478.
- Landan, G., A.Bdolah, Z.Wollberg, E.Kochva, and D.Graur. 1991a. "Evolution of the sarafotoxin/endothelin superfamily of proteins." *Toxicon.* 29:237-244.
- Landan, G., A.Bdolah, Z.Wollberg, E.Kochva, and D.Graur. 1991b. "Evolution of the sarafotoxin/endothelin superfamily of proteins." *Toxicon.* 29:237-244.
- Liebich, I., J.Bode, I.Reuter, and E.Wingender. 2002. "Evaluation of sequence motifs found in scaffold/matrix-attached regions (S/MARs)." *Nucleic Acids Res.* 30:3433-3442.
- Lynch, V.J. 2007. "Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A2 genes." *BMC.Evol.Biol.* 7:2.
- Ma, D., A.Armugam, and K.Jeyaseelan. 2001. "Expression of cardiotoxin-2 gene. Cloning, characterization and deletion analysis of the promoter." *Eur.J.Biochem.* 268:1844-1850.
- Ma, D., A.Armugam, and K.Jeyaseelan. 2002. "Alpha-neurotoxin gene expression in *Naja sputatrix*: identification of a silencer element in the promoter region." *Arch.Biochem.Biophys.* 404:98-105.

- Mann, K.G., M.F.Hockin, K.J.Begin, and M.Kalafatis. 1997. "Activated protein C cleavage of factor Va leads to dissociation of the A2 domain." *J.Biol.Chem.* 272:20678-20683.
- Markland, F.S. 1997. "Snake venoms." *Drugs.* 54 Suppl 3:1-10.
- Markland, F.S. 1998. "Snake venoms and the hemostatic system." *Toxicon.* 36:1749-1800.
- Marsh, N.A., T.L.Fyffe, and E.A.Bennett. 1997. "Isolation and partial characterization of a prothrombin-activating enzyme from the venom of the Australian rough-scaled snake (*Tropidechis carinatus*)." *Toxicon.* 35:563-571.
- Masci, P.P., A.N.Whitaker, and J.de Jersey. 1988. "Purification and characterization of a prothrombin activator from the venom of the Australian brown snake, *Pseudonaja textilis textilis*." *Biochem.Int.* 17:825-835.
- McMullen, B.A., K.Fujikawa, W.Kisiel, T.Sasagawa, W.N.Howald, E.Y.Kwa, and B.Weinstein. 1983. "Complete amino acid sequence of the light chain of human blood coagulation factor X: evidence for identification of residue 63 as beta-hydroxyaspartic acid." *Biochemistry.* 22:2875-2884.
- Minh, L.T., M.A.Reza, S.Swarup, and R.M.Kini. 2005. "Gene duplication of coagulation factor V and origin of venom prothrombin activator in *Pseudonaja textilis* snake." *Thromb.Haemost.* 93:420-429.
- Morita, T. 2004. "C-type lectin-related proteins from snake venoms." *Curr.Drug Targets.Cardiovasc.Haematol.Disord.* 4:357-373.
- Morita, T. and S.Iwanaga. 1978. "Purification and properties of prothrombin activator from the venom of *Echis carinatus*." *J.Biochem.(Tokyo).* 83:559-570.
- Moura-da-Silva, A.M., R.D.Theakston, and J.M.Crampton. 1996. "Evolution of disintegrin cysteine-rich and mammalian matrix-degrading metalloproteinases: gene duplication and divergence of a common ancestor rather than convergent evolution." *J.Mol.Evol.* 43:263-269.
- Nesheim, M.E., W.M.Canfield, W.Kisiel, and K.G.Mann. 1982. "Studies of the capacity of factor Xa to protect factor Va from inactivation by activated protein C." *J.Biol.Chem.* 257:1443-1447.
- Nesheim, M.E., J.B.Taswell, and K.G.Mann. 1979. "The contribution of bovine Factor V and Factor Va to the activity of prothrombinase." *J.Biol.Chem.* 254:10952-10962.
- Nishida, S., T.Fujita, N.Kohno, H.Atoda, T.Morita, H.Takeya, I.Kido, M.J.Paine, S.Kawabata, and S.Iwanaga. 1995. "cDNA cloning and deduced amino acid sequence of prothrombin activator (ecarin) from Kenyan *Echis carinatus* venom." *Biochemistry.* 34:1771-1778.
- Owen, W.G. and C.M.Jackson. 1973. "Activation of prothrombin with *Oxyranus scutellatus* scutellatus (Taipain snake) venom." *Thromb.Res.* 3:705-714.
- Pittman, D.D., K.N.Tomkinson, D.Michnick, U.Selighsohn, and R.J.Kaufman. 1994. "Posttranslational sulfation of factor V is required for efficient thrombin cleavage and activation and for full procoagulant activity." *Biochemistry.* 33:6952-6959.
- Rao, V.S., J.S.Joseph, and R.M.Kini. 2003a. "Group D prothrombin activators from snake venom are structural homologues of mammalian blood coagulation factor Xa." *Biochem.J.* 369:635-642.

- Rao, V.S. and R.M.Kini. 2002. "Pseutarin C, a prothrombin activator from *Pseudonaja textilis* venom: its structural and functional similarity to mammalian coagulation factor Xa-Va complex." *Thromb.Haemost.* 88:611-619.
- Rao, V.S., S.Swarup, and R.M.Kini. 2003b. "The nonenzymatic subunit of pseutarin C, a prothrombin activator from eastern brown snake (*Pseudonaja textilis*) venom, shows structural similarity to mammalian coagulation factor V." *Blood.* 102:1347-1354.
- Rao, V.S., S.Swarup, and R.M.Kini. 2004. "The catalytic subunit of pseutarin C, a group C prothrombin activator from the venom of *Pseudonaja textilis*, is structurally similar to mammalian blood coagulation factor Xa." *Thromb.Haemost.* 92:509-521.
- Reza, A., S.Swarup, and R.M.Kini. 2005a. "Two parallel prothrombin activator systems in Australian rough-scaled snake, *Tropidechis carinatus*. Structural comparison of venom prothrombin activator with blood coagulation factor X." *Thromb.Haemost.* 93:40-47.
- Reza, M.A., L.T.Minh, S.Swarup, and R.M.Kini. 2006. "Molecular evolution caught in action: gene duplication and evolution of molecular isoforms of prothrombin activators in *Pseudonaja textilis* (brown snake)." *J.Thromb.Haemost.* 4:1346-1353.
- Reza, M.A., S.Swarup, and R.M.Kini. 2005b. "Gene structures of trocarin D and coagulation factor X, two functionally diverse prothrombin activators from Australian rough scaled snake." *Pathophysiol.Haemost.Thromb.* 34:205-208.
- Reza, M.A., S.Swarup, and R.M.Kini. 2007. "Structure of two genes encoding parallel prothrombin activators in *Tropidechis carinatus* snake: gene duplication and recruitment of factor X gene to the venom gland." *J.Thromb.Haemost.* 5:117-126.
- Robin Doley, Xingding Zhou, and R.M.Kini. 2009. "Snake Venom Phospholipase A2 Enzymes." In Stephen P.Mackessy, editor, *Handbook of Venoms and Toxins of Reptiles*. CRC Press. 173-205.
- Rosing, J. and G.Tans. 1991. "Inventory of exogenous prothrombin activators. For the Subcommittee on Nomenclature of Exogenous Hemostatic Factors of the Scientific and Standardization Committee of the International Society on Thrombosis and Haemostasis." *Thromb.Haemost.* 65:627-630.
- Rosing, J. and G.Tans. 1992. "Structural and functional properties of snake venom prothrombin activators." *Toxicon.* 30:1515-1527.
- Schieck, A., E.Habermann, and F.Kornalik. 1972. "The prothrombin-activating principle from *Echis carinatus* venom. II. Coagulation studies in vitro and in vivo." *Naunyn Schmiedebergs Arch.Pharmacol.* 274:7-17.
- Silva, M.B., M.Schattner, C.R.Ramos, I.L.Junqueira-de-Azevedo, M.C.Guarnieri, M.A.Lazzari, C.A.Sampaio, R.G.Pozner, J.S.Ventura, P.L.Ho, and A.M.Chudzinski-Tavassi. 2003. "A prothrombin activator from *Bothrops erythromelas* (jararaca-da-seca) snake venom: characterization and molecular cloning." *Biochem.J.* 369:129-139.
- Snow, C.M., A.Senior, and L.Gerace. 1987. "Monoclonal antibodies identify a group of nuclear pore complex glycoproteins." *J.Cell Biol.* 104:1143-1156.

- Speijer, H., J.W.Govers-Riemslog, R.F.Zwaal, and J.Rosing. 1986. "Prothrombin activation by an activator from the venom of *Oxyuranus scutellatus* (Taipan snake)." *J.Biol.Chem.* 261:13258-13267.
- St Pierre, L., S.T.Earl, I.Filippovich, N.Sorokina, P.P.Masci, J.De Jersey, and M.F.Lavin. 2008. "Common evolution of waprin and kunitz-like toxin families in Australian venomous snakes." *Cell Mol.Life Sci.* 65:4039-4054.
- St Pierre, L., P.P.Masci, I.Filippovich, N.Sorokina, N.Marsh, D.J.Miller, and M.F.Lavin. 2005. "Comparative analysis of prothrombin activators from the venom of Australian elapids." *Mol.Biol.Evol.* 22:1853-1864.
- Stenflo, J., A.Lundwall, and B.Dahlback. 1987. "beta-Hydroxyasparagine in domains homologous to the epidermal growth factor precursor in vitamin K-dependent protein S." *Proc.Natl.Acad.Sci.U.S.A.* 84:368-372.
- Stocker, K., H.Hauer, C.Muller, and D.A.Triplett. 1994. "Isolation and characterization of Textarin, a prothrombin activator from eastern brown snake (*Pseudonaja textilis*) venom." *Toxicon.* 32:1227-1236.
- Suzuki, K., B.Dahlback, and J.Stenflo. 1982. "Thrombin-catalyzed activation of human coagulation factor V." *J Biol Chem.* 257:6556-6564.
- Tans, G., J.W.Govers-Riemslog, J.L.van Rijn, and J.Rosing. 1985. "Purification and properties of a prothrombin activator from the venom of *Notechis scutatus scutatus*." *J.Biol.Chem.* 260:9366-9372.
- Thorelli, E., R.J.Kaufman, and B.Dahlback. 1998. "Cleavage requirements of factor V in tissue-factor induced thrombin generation." *Thromb.Haemost.* 80:92-98.
- van der Neut, K.M., R.J.Dirven, H.L.Vos, G.Tans, J.Rosing, and R.M.Bertina. 2004b. "Factor Va is inactivated by activated protein C in the absence of cleavage sites at Arg-306, Arg-506, and Arg-679." *J.Biol.Chem.* 279:6567-6575.
- van der Neut, K.M., R.J.Dirven, H.L.Vos, G.Tans, J.Rosing, and R.M.Bertina. 2004a. "Factor Va is inactivated by activated protein C in the absence of cleavage sites at Arg-306, Arg-506, and Arg-679." *J.Biol.Chem.* 279:6567-6575.
- van Drunen, C.M., R.W.Oosterling, G.M.Keultjes, P.J.Weisbeek, R.van Driel, and S.C.Smeekens. 1997. "Analysis of the chromatin domain organisation around the plastocyanin gene reveals an MAR-specific sequence element in *Arabidopsis thaliana*." *Nucleic Acids Res.* 25:3904-3911.
- Walker, F.J., W.G.Owen, and C.T.Esmon. 1980. "Characterization of the prothrombin activator from the venom of *Oxyuranus scutellatus scutellatus* (taipan venom)." *Biochemistry.* 19:1020-1023.
- Wang, C., M.Eufemi, C.Turano, and A.Giartosio. 1996. "Influence of the carbohydrate moiety on the stability of glycoproteins." *Biochemistry.* 35:7299-7307.
- Welton, R.E. and J.N.Burnell. 2005. "Full length nucleotide sequence of a factor V-like subunit of oscurarin from *Oxyuranus scutellatus scutellatus* (coastal Taipan)." *Toxicon.* 46:328-336.
- Wilberding, J.A. and F.J.Castellino. 2000. "Characterization of the murine coagulation factor X promoter." *Thromb.Haemost.* 84:1031-1038.

- Will, K., G.Warnecke, N.Albrechtsen, T.Boulikas, and W.Deppert. 1998. "High affinity MAR-DNA binding is a common property of murine and human mutant p53." *J.Cell Biochem.* 69:260-270.
- Yamada, D. and T.Morita. 1997. "Purification and characterization of a Ca<sup>2+</sup> -dependent prothrombin activator, multactivase, from the venom of *Echis multisquamatus*." *J.Biochem.(Tokyo)*. 122:991-997.
- Zhou, C.Z. and B.Liu. 2001. "Identification and characterization of a silkgland-related matrix association region in *Bombyx mori*." *Gene*. 277:139-144.

# A Puroindoline Mutigene Family Exhibits Sequence Diversity in Wheat and is Associated with Yield-Related Traits

Feng Chen<sup>1,2</sup>, Fuyan Zhang<sup>1</sup>,  
Craig F. Morris<sup>3</sup> and Dangqun Cui<sup>1,2</sup>

<sup>1</sup>Department of Agronomy, Henan Agricultural University, Zhengzhou

<sup>2</sup>Key Laboratory of Physiological Ecology and Genetic Improvement  
of Food Crops in Henan Province, Zhengzhou

<sup>3</sup>USDA-ARS, Western Wheat Quality Laboratory, E-202 Food Science and Human  
Nutrition Facility East, Washington State University, Pullman,

<sup>1,2</sup>China

<sup>3</sup>USA

## 1. Introduction

Kernel texture (grain hardness) is a leading quality characteristic of bread wheat (*Triticum aestivum* L.) as it dramatically influences the milling and processing properties, and consequently determines the classification and marketing of grain (Bhave and Morris 2008a, b). The word puroindoline is derived from the Greek word “*puros*” meaning wheat and “*indoline*” describing the indole ring of tryptophan (Gautier et al. 1994). Puroindolines, composed of puroindoline a and b, are amphipathic proteins of ca. 13,000 Da, and share homology with grain softness protein (GSP), purothionins, lipid transfer proteins, and other members of the prolamin super-family of proteins (Shewry and Halford 2002). Puroindoline proteins possess a characteristic tryptophan-rich domain and cysteine backbone; isoforms occur in the starchy endosperm of the Triticeae. Their secondary structure, determined by infrared and Raman spectroscopies, is comprised of approximately 30%  $\alpha$ -helices, 30%  $\beta$ -sheets, and 40% unordered structure at pH 7.4 in solution (Bihan et al. 1996). Puroindoline genes are present throughout the Triticeae tribe of the Poaceae (Gramineae), including wheat (*Triticum* sp.), rye (*Secale* sp.), barley (*Hordeum* sp.), and the wild relatives of wheat (*Aegilops* sp. and *Triticum* sp.). In *Triticum aestivum*, puroindolines exist as two expressed genes, Puroindoline a and Puroindoline b, and are located on the distal end of the short arm of chromosome 5 (5DS). An exception to this general situation lies with the tetraploid (AABB) wheats (*T. turgidum*), which include cultivated durum (ssp. *durum*). Apparently during the allotetraploidization formation of *T. diccoides*, the wild ancestor of cultivated durum, both the A- and B-genome Puroindoline loci were eliminated due to transposable element insertion and two large deletions in the *Hardness* loci caused by illegitimate DNA recombination (Chantret et al. 2005). Consequently, hexaploid wheat, *T. aestivum*, possesses

puroindoline a and b on only the D-genome (contributed by *Ae. tauschii* during the allohexaploidization of this taxon), but lacks homoeologous loci on the A- and B-genomes. The expression of puroindoline genes is mainly associated with 'soft' and 'hard' grain texture of bread wheat, whereas durum wheat is 'very hard' due to the lack of puroindoline genes (Capparelli et al. 2003). Extensive genetic surveys, cytological analysis, and transformation experiments in wheat and rice (*Oryza sativa*) have demonstrated that puroindoline a and b act to create soft kernel texture (Bhave and Morris 2008a, b, Morris 2002, Morris and Bhave 2008). However, puroindolines have also been extensively studied regarding their multiple roles in the development and resistance of plants against biotic stress, and their profound influence on the breeding and processing quality of food crops (Luo et al. 2008; Giroux and Morris 1997, 1998; Ragupathy and Cloutier 2008; Chen et al. 2007a, b). Considerable allelic variation in puroindoline a and b have been identified in a great number of genotypes stemming from different geographies. Gene sequence polymorphism and allele designations have been recently reconciled and reviewed by Morris and Bhave (2008) and Bhave and Morris (2008a, b).

Although clearly exerting a major role in wheat kernel texture variation, puroindolines do not account for all of the variation observed in bread wheat. Several minor quantitative trait loci (QTLs) for kernel texture have been mapped on different chromosomes (Sourdille et al. 1996; Turner et al. 2004). Four additional regions located on chromosomes 2A, 2D, 5B, and 6D were shown to have single-factor effects on hardness, while three others located on chromosomes 5A, 6D and 7A had interaction effects (Sourdille et al. 1996).

Three new puroindoline gene-like sequences in hexaploid wheat were reported by Wilkinson et al. (2008). Chen et al. (2010) renamed them *Pinb-2v1*, *Pinb-2v2* and *Pinb-2v3*, and reported the discovery of a fourth novel puroindoline variant in bread wheat, designated *Pinb-2v4*. Physical mapping results of Chen et al. (2010) were not fully consistent with the results reported by Wilkinson et al. (2008). Interestingly, *Pinb-2v1* and *Pinb-2v4* were present in all of the surveyed cultivars whereas *Pinb-2v2* and *Pinb-2v3* were reciprocally present in different wheat cultivars (Chen et al. 2010). This result indicates that *Pinb-2v2* and *Pinb-2v3* are likely allelic. In this study, we remapped the four known *Puroindoline b-2* variants using aneuploids of bread wheat and durum wheat in order to confirm their physical location on the chromosomes of wheat, and report the discovery of a novel *Pinb-2v3* allele and a new *Puroindoline b-2* variant, designated *Pinb-2v5*. Also, we investigated the association of *Puroindoline b-2* variants with yield-related traits. This study provides useful information for illustrating the molecular and genetic basis of kernel hardness and gene duplication events in wheat.

## 2. Materials and methods

### 2.1 Wheat germplasm, DNA extraction and PCR amplification

A total of 109 bread wheat lines developed for the Yellow and Huai Valleys of China were planted at the Zhengzhou Scientific Research and Education Center of Henan Agricultural University during the 2009-10 cropping seasons according to local management practices. Analysis of agronomic traits and measurement of SKCS hardness as well as identification of puroindoline b-2 variants were performed. Each plot was comprised of four 200 cm-long rows with 23 cm between neighboring rows and 10 cm between neighboring plants. All surveyed cultivars grew well and no lodging was present in the trial. After harvest, all wheat samples were cleaned.



Durum wheat cultivar Langdon (LND), bread wheat cultivar Chinese Spring (CS), two disomic substitution lines of Langdon 7D(7A) and Langdon 7D(7B) with substituted 7D chromosomes of Chinese Spring and six nullisomic-tetrasomic lines of Chinese Spring involving group 7 chromosomes (CS N7A-T7B, CS N7A-T7D, CS N7B-T7A, CS N7B-T7D, CS N7D-T7A and CS N7D-T7B) were used in this study. DNA was extracted from two seeds each of the 109 bread wheat lines, genetic stocks, and 15 Chinese, CIMMYT and Italian durum wheat cultivars following the rapid extraction method of genomic DNA derived from Chen et al. (2006).

PCR primer sequences were designed using Primer Premier 5.0 software. Reactions were performed in 25  $\mu$ L containing 100 ng of genomic DNA, 10 pmol of each primer, 250  $\mu$ M of each dNTP, 1x Taq DNA polymerase reaction buffer containing 1.5  $\mu$ M of  $MgCl_2$  and 0.5 unit of Taq DNA polymerase (Promega, Madison, WI). The cycling conditions were 94°C for 5 min followed by 35 cycles of 94°C for 50 s, 45°C to 65°C for 50 s (primer-specific annealing temperatures, see Table 1), 72°C for 1 min, followed by a final 10-min extension at 72°C. An aliquot (8  $\mu$ L) of the PCR products was analyzed on 1.5% (w/v) agarose gels, stained with ethidium bromide, and visualized with UV light. Forty-eight clones from three independent PCR reactions using the universal primers on durum cv. Langdon were sequenced from both strands by SinoGenoMax Co., Ltd.

Multiple alignments of sequences and translations of nucleotide into amino acid sequence were performed by software DNAMAN Version 6.0. Sequence chromatograms were analyzed by Chromas Version 1.4.5 and FinchTV Version 1.4.0.

## 2.2 Measurement of agronomic traits and hardness index

Before harvesting, ten plants of each genotype were randomly selected in each plot to determine spikelet number per plant, flag leaf length and flag leaf width, as well as flag leaf area calculated from the product of leaf length and maximum leaf width  $\times$  0.75. Grains per spike, grain weight per plant and grain weight per spike were determined after harvesting. Kernel hardness, grain diameter and thousand-kernel weight were measured on 300-kernel samples of each genotype using the Perten Single Kernel Characterization System (SKCS) 4100, following the manufacturer's operation procedure (Perten Instruments North America Inc., Springfield, IL). Mean, standard deviation, and distribution of SKCS hardness data were used to classify the genotypes tested into soft, mixed, and hard types.

## 3. Results

### 3.1 Physical mapping of *Pinb-2* variants

Physical mapping results with variant-specific primers (Table 1) showed that *Pinb-2v1* was present in LND 7D(7A), LND 7D(7B), Chinese Spring, CS N7A-T7B, CS N7A-T7D, CS N7B-T7A and CS N7B-T7D, but absent in wild-type Langdon, CS N7D-T7A and CS N7D-T7B, indicating that *Pinb-2v1* is located on 7D of Chinese Spring. *Pinb-2v2* was present in wild-type Chinese Spring, CS N7A-T7B, CS N7A-T7D, CS N7D-T7A and CS N7D-T7B, but absent in wild-type Langdon, LND 7D(7A), LND 7D(7B), CS N7B-T7A and CS N7B-T7D, indicating that *Pinb-2v2* is located on 7B of Chinese Spring. *Pinb-2v3* was present in wild-type Langdon, LND 7D(7A), but absent in LND 7D(7B), wild-type Chinese Spring and all of its six nullisomic-tetrasomic lines, indicating that *Pinb-2v3* is located on 7B of Langdon. *Pinb-2v4* was present in Langdon, LND 7D(7B), CS N7B-T7A, CS N7B-T7D, CS N7D-T7A and CS

N7D-T7B, but absent in LND 7D(7A), CS N7A-T7B and CS N7A-T7D, indicating that *Pinb-2v4* is located on 7A of Langdon and Chinese Spring.

Therefore, it is concluded that *Pinb-2v1* is on 7D of Chinese Spring, *Pinb-2v2* is on 7B of Chinese Spring, *Pinb-2v3* is on 7B of Langdon and *Pinb-2v4* is on 7A of Langdon and Chinese Spring. The results of mapping *Pinb-2v* in Chinese Spring and its derived nullisomic-tetrasomic lines are consistent with the locations of the four *Puroindoline b-2* variants reported by Chen et al. (2010a).

### 3.2 Variation in *puroindoline b-2* variants and the discovery of a new *puroindoline b-2* variant

Based on PCR amplification with four specific primers (Table 1), 35 of 36 durum wheat cultivars surveyed possessed variant combination *Pinb-2v3/Pinb-2v4*, whereas the durum wheat cultivar Norba, introduced from Italy, possessed combination *Pinb-2v2/Pinb-2v4*. None of the surveyed durum cultivars possessed the *Pinb-2v1* gene. Further sequencing of PCR products with specific primers showed that all of the sequences of *Pinb-2v4* were exactly the same as the *Pinb-2v4* sequence of bread wheat reported by Chen et al. (2010), whereas the coding region of *Pinb-2v3* in durum wheat cultivar Langdon had a single nucleotide change from G to T at the 6th position compared with the *Pinb-2v3* sequence reported by Wilkinson et al. (2008) and Chen et al. (2010). Furthermore, 18 durum wheat cultivars with the *Pinb-2v3* variant were sequenced and all of them uniformly contained a single nucleotide change from G to T at the 6th position. Sequencing of bread wheat cultivar Wichita showed the same change at the 6th position, compared with a previous *Pinb-2v3* sequence of Wichita reported by Wilkinson et al. (2008) and Chen et al. (2010). The discrepancy may have resulted from the previously used primers due to a single nucleotide change present in the forward primer sequence. However, after we modified the *Pinb-2v3* sequence from G to T at the 6th position of Wichita, a one base-pair change from T to C at the 311th position was found in the coding region of *Pinb-2v3* in 12 durum wheat cultivars, resulting in the deduced amino acid change from valine to alanine at position 104 in PINB-2v3. According to the gene nomenclature for *puroindoline a* and *b*, this *Pinb-2v3* allele in wheat cultivars containing the sequence of AM99733 and GQ496618 with the modification of G to T at the 6th position will be designated as *Pinb-2v3a*, and the *Pinb-2v3* allele with a nucleotide change from T to C at the 311th position in durum wheat cultivars will be designated as *Pinb-2v3b*.

Three repeated PCR amplifications of the durum cultivar Langdon were performed with the universal primer in order to reduce the influence of mismatching amplification of *Taq* polymerase. Cloning and sequencing results showed that 40 sub-clones belonged to either *Pinb-2v3* or *Pinb-2v4* whereas 7 other sub-clones did not belong to any known *Puroindoline b-2* variant. Further analysis indicated that a novel variant, designated as *Pinb-2v5* (HM245236), shared high homology and was most closely related to *Pinb-2v4* with only 5 SNPs (Fig. 1).

The overall alignment with *Pinb-D1a* with the five *Pinb-2* variant genes showed that *Pinb-2v5* has 74.3%, 95.8%, 94.0%, 92.3% and 98.8% identity with *Pinb-D1a*, *Pinb-2v1*, *Pinb-2v2*, *Pinb-2v3* and *Pinb-2v4*, respectively, at the DNA level. Moreover, all four *puroindoline b* variant sequences were at least 91.8% homologous amongst themselves. Full alignment of deduced amino acid sequences of *Pinb-D1a* and the five *Puroindoline b-2* variant genes indicated that

*Pinb-2v5* has 57.8%, 96.0%, 91.3%, 87.9% and 96.5% identity with *Pinb-D1a*, *Pinb-2v1*, *Pinb-2v2*, *Pinb-2v3* and *Pinb-2v4*, respectively, at the amino acid level (data not shown).

Gene	Forward primer	Reverse primer	PCR annealing Temp °C	Fragment size (bp)
<i>Pinb-2vU</i>	ATGAAGACCTTATTCCTCCTA	TCASTAGTAATAGCCATTAKT A	54	453
<i>Pinb-2v1</i>	GGTTCTCAAACTGCCCCAT	ACTTGCAGTTGGAATCCAG	57	319
<i>Pinb-2v2</i>	CTTGTAAGTACACAACTTT GCA	GTATGGACGAACTTGCAGCTG GAG	65	401
<i>Pinb-2v3</i>	GCAGAAAAAGCCATTGCACCT A	CATTAGTAGGGACGAACTTGC AGCTA	65	528
<i>Pinb-2v4</i>	CCTTTCTCTTGTAGTGAGCAC AACCA	GACGAACTTGCAGTTGGAATC CAA	65	403
<i>Pinb-D1b</i>	ATGAAGGCCCTCTTCCTCA	CTCATGCTCACAGCCGCT	58	250
<i>N-Pinb-D1b</i>	ATGAAGGCCCTCTTCCTCA	CTCATGCTCACAGCCGCC	58	250

Table 1. PCR primers used in generating *Puroindoline b-2* variant gene sequences in durum wheat

### 3.3 Association of puroindoline b-B2 variants with SKCS grain hardness, other grain traits, grain yield components, and flag leaf size

In order to investigate the influence of *puroindoline b-2* variants on kernel hardness independent of puroindoline a and b alleles, the average SKCS hardness index of two different *Puroindoline b-2* variant combinations, *Pinb-2v1/Pinb-2v2/Pinb-2v4* and *Pinb-2v1/Pinb-2v3/Pinb-2v4*, were compared by sub-setting the lines according to *Pin-D1* haplotype. Puroindoline alleles are associated with dramatic effects on kernel hardness (reviewed in Morris, 2002; Morris and Behave, 2008; Behave and Morris, 2008a, b). Results indicated that in soft wheat cultivars with *Pina-D1a/Pinb-D1a*, the average of SKCS hardness index of cultivars with *Pinb-2v3* was 24.6, which was significantly higher than that of cultivars with *Pinb-2v2* (14.2). No significant difference was observed among cultivars with *Pinb-D1b*, possibly suggesting that the *Puroindoline b-2* variants had a larger impact in soft wheat than in hard.

No significant association between the puroindoline D1 alleles and the 9 agronomic traits relating to wheat yield were observed in the surveyed Chinese wheat cultivars of the Yellow and Huai Valleys, suggesting that the puroindoline genes have no apparent relationship with wheat grain yield. However, when the agronomic traits of cultivars with different *Puroindoline b-2* variant combinations were compared, the averages of grain number per spike, grain weight per spike, width of flag leaf and flag leaf area of cultivars with *Pinb-2v3* were significantly higher than those cultivars with *Pinb-2v2*. Differences in thousand kernel weight, grain length and spikelet number per spike were not different between these two *Pinb-2v2* and

*Pinb-2v3* genotypes (Table 2). The results indicated that *Pinb-2v3* cultivars possessed superior grain yield traits compared to *Pinb-2v2*, possibly suggesting that the *Puroindoline b-2* gene could modulate wheat grain yield to some extent, and that the grain yield of cultivars with *Pinb-2v3* is slightly higher than that of cultivars with *Pinb-2v2*. Another likely possibility is that these *Pinb-2* variants are identifying germplasm pools or founder effects.

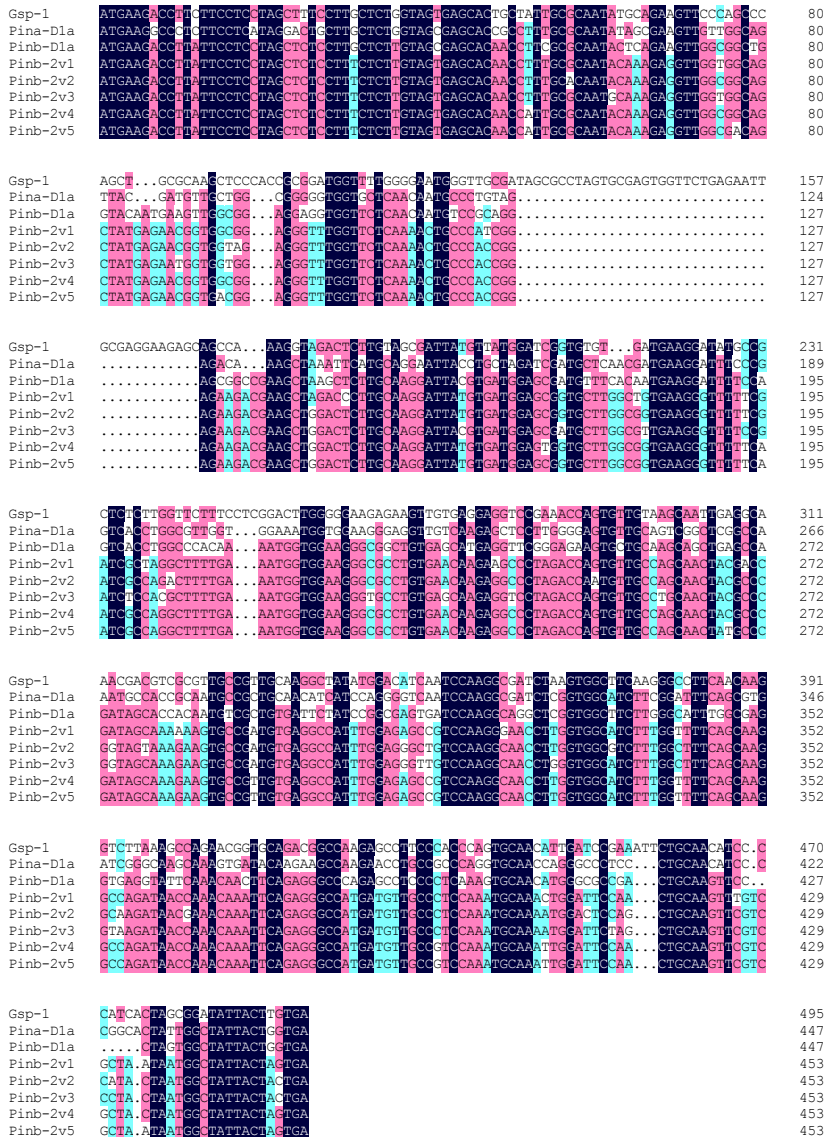


Fig. 1. DNA sequence alignment of *Puroindoline b-2* gene variants *Pinb-2v1*, *Pinb-2v2*, *Pinb-2v3*, *Pinb-2v4*, *Pinb-2v5*, *Pinb-D1a*, *Pina-D1a* as well as *Gsp-1* from bread wheat.

	Wild type		<i>Pinb-D1b</i>		Total	
	<i>Pinb-2v2</i>	<i>Pinb-2v3</i>	<i>Pinb-2v2</i>	<i>Pinb-2v3</i>	<i>Pinb-2v2</i>	<i>Pinb-2v3</i>
No. of sample	2	23	19	48	21	71
SKCS Hardness	14.2a	24.6b	67.5a	60.7b	-	-
1000-kernel weight	59.2a	54.7b	50.9a	52.2b	51.7a	52.9a
Grain length (mm)	6.85a	6.8a	6.61a	6.73a	6.63a	6.66a
Grain width (mm)	3.65a	3.57a	3.47a	3.55a	3.49a	3.50a
Spikelet number per spike	20.1a	20.83a	20.09a	20.59a	20.10a	20.67a
Grain number per spike	42.1a	45.18b	40.7a	43.7b	40.86a	44.17b
Grain weight per spike (g)	2.11a	2.27a	1.95a	2.17b	1.96a	2.21b
Length of flag leaf (cm)	16.9a	17.4a	15.7a	17.9b	15.77a	17.70b
Width of flag leaf (cm)	1.85a	1.96a	1.85a	1.92a	1.85a	1.93a
Area of flag leaf (cm <sup>2</sup> )	23.6a	25.6b	21.7a	25.7b	21.89a	25.69b

Different letters indicate significant difference at 5% probability level

Table 2. Comparison of two *Puroindoline b*-B2 variants on grain hardness and yield-related traits

#### 4. Discussion

Even though physical mapping of *Puroindoline b*-2 variants was conducted using aneuploids of Chinese Spring by Chen et al. (2010), we re-mapped the *Puroindoline b*-2 variants in the durum wheat cultivar Langdon and bread wheat cultivar Chinese Spring. The re-mapping was performed using different stocks than those reported by Chen et al. (2010) due to the controversial difference between the results of Chen et al. (2010) and Wilkinson et al. (2008). According to this study and the previous report (Chen et al. 2010), *Pinb-2v1* is located on chromosome 7D and is present in all the bread wheat cultivars surveyed. Therefore, in this study, the absence of *Pinb-2v1* in durum wheat is expected. *Pinb-2v2* and *Pinb-2v3* are reciprocally present on chromosome 7B in all of the bread wheat cultivars surveyed. Only one durum cultivar possessed the *Pinb-2v2/Pinb-2v4* haplotype combination, whereas 35 durum cultivars possessed the *Pinb-2v3/Pinb-2v4* combination, suggesting that *Pinb-2v3* and *Pinb-2v2* were likely allelic. In this set of durum germplasm, the *Pinb-2v3/Pinb-2v4* combination was the predominant haplotype.

Wilkinson et al. (2008) mapped *Puroindoline b*-2 variant 2 to chromosome 7AL in three doubled haploid populations and amplified sequences of *Pinb-2v3* and *Pinb-2v1* from the genomic DNA of the durum wheat cultivar Ofanto. We did not find *Pinb-2v1* in any of the durum cultivars surveyed, including 48 sequencing results of cloned PCR amplicon in the

durum wheat Langdon. The reason for this discrepancy is possibly due to primer sequence and specificity (or lack thereof) for the various variant genes used in the two studies.

According to Chen et al. (2010), *Pinb-2v1*, *Pinb-2v2*, *Pinb-2v3* and *Pinb-2v4* are located on chromosome 7DL, 7BL, 7B and 7AL in bread wheat, respectively. Coincidentally, the strongest QTL effects controlling grain yield, especially for grain weight, were found on chromosomes 7AL and 7BL in the report of Quarrie et al. (2005), and a QTL associated with grain yield has also been identified on chromosome 7D in the study of Kuchel et al. (2007). More recently, several QTLs including QTLs on 7A and 7B associated with grain yield and yield components have been discovered in a recombinant inbred line population (McIntyre et al. 2009). However, further studies are required to determine if this is 'cause and effect' or simply linkage occurring.

Even though many QTLs controlling wheat grain yield and its components have been studied for many years, no specific gene with detailed sequence has been reported so far. Therefore, the possibility of the *Puroindoline b-2* gene possessing some function in modulating grain yield traits may provide useful information for MAS (Marker Assisted Selection). Future studies should define the function of the *Puroindoline b-2* genes by using populations with defined genetic background.

## 5. Acknowledgements

This project was funded by the National Natural Science Foundation (31000708), Henan Provincial Key Technologies R & D Program (114300510013) and Specialized Research Fund for the Doctoral Program of Higher Education (20104105120003) of China.

## 6. References

- Bhave M, Morris CF. 2008a. Molecular genetics of puroindolines and related genes: allelic diversity in wheat and other grasses. *Plant Molecular Biology* 66, 205-219.
- Bhave M Morris CF. 2008b. Molecular genetics of puroindolines and related genes: regulation of expression, membrane binding properties and applications. *Plant Molecular Biology* 66, 221-231.
- Bihan TL, Blochet JE, Desormeaux A, Marion D, Pezolet M. 1996. Determination of the secondary structure and conformation of puroindolines by infrared and Raman spectroscopy. *Biochemistry* 35: 12712-12722.
- Capparelli R, Borriello G, Giroux MJ, Amoroso MG. 2003. Puroindoline a-gene expression is involved in association of puroindolines to starch. *Theoretical and Applied Genetics* 107: 1463-1468.
- Chantret N, Salse J, Sabot F, Rahman S, Bellec A, Laubin B, Dubois I, Dossat C, Sourdille P, Joudrier P, Gautier MF, Cattolico L, Beckert M, Aubourg S, Weissenbach J, Caboche M, Bernard M, Leroy P, Chalhou B. 2005. Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell* 17: 1033-1045.
- Chen F, Beecher B, Morris CF. 2010a. Physical mapping and a new variant of *Puroindoline b-2* genes in wheat. *Theoretical and Applied Genetics* 120:745-751.

- Chen F, He ZH, Chen DS, Zhang CL, Zhang Y, Xia XC. 2007a. Influence of puroindoline alleles on milling performance and qualities of Chinese noodles, steamed bread and pan bread in spring wheats. *Journal of Cereal Science* 45: 59-66.
- Chen F, He ZH, Xia XC, Xia LQ, Zhang XY, Lillemo M, Morris CF. 2006. Molecular and biochemical characterization of puroindoline a and b alleles in Chinese landraces and historical cultivars. *Theoretical and Applied Genetics* 112: 400-409.
- Chen F, Yu Y, Xia X, He Z. 2007b. Prevalence of a novel puroindoline b allele in Yunnan endemic wheats (*Triticum aestivum* ssp. *Yunnanense* King). *Euphytica* 156: 39-46.
- Giroux MJ, Morris CF. 1997. A glycine to serine change in puroindoline b is associated with wheat grain hardness and low levels of starch-surface friabilin. *Theoretical and Applied Genetics* 95, 857-864.
- Giroux MJ, Morris CF. 1998. Wheat grain hardness results from highly conserved mutations in the friabilin components puroindoline a and b. *Proceedings of the National Academy of the United States of America* 95, 6262-6266.
- Kuchel H, Williams JK., Langridge P, Eagles HA, Jefferies SP. 2007. Genetic dissection of grain yield in bread wheat. I. QTL analysis, *Theoretical and Applied Genetics* 115: 1029-1041.
- Luo LT, Zhang JR, Yang GX, Li Y, Li KX, He GY. 2008. Expression of puroindoline a enhances leaf rust resistance in transgenic tetraploid wheat. *Molecular Biology Reports* 35: 195-200.
- McIntyre CL, Mathews KL, Rattey A, Chapman SC, Drenth J, Ghaderi, M, Reynolds M, Shorter R. 2010. Molecular detection of genomic regions associated with grain yield and yield-related components in an elite bread wheat cross evaluated under irrigated and rainfed conditions. *Theoretical and Applied Genetics*, 120: 527-541.
- Morris CF. 2002. Puroindolines, the molecular genetic basis of wheat grain hardness. *Plant Molecular Biology* 48, 633-647.
- Morris CF, Bhavé M. 2008. Reconciliation of D-genome puroindoline allele designations with current DNA sequence data. *Journal of Cereal Science* 48, 277-287.
- Quarrie SA, Steed A, Calestani C, Semikbodskii A, Lebreton C, Chinoy C, Steele N, Pljevljakusic D, Habash DZ, Farmer P, Saker L, Clarkson DT, Abugalieva A, Yessimbekova M, Turuspekov Y, Abugalieva S, Tuberosa R, Sanguineti M-C, Hollington PA, Aragues R, Royo A, Dodig D. 2005. A high density genetic map of hexaploid wheat (*Triticum aestivum* L.) from the cross Chinese Spring x SQ1 and its use to compare QTLs for grain yield across a range of environments. *Theoretical and Applied Genetics* 110: 865-880.
- Ragupathy R, Cloutier S. 2008. Genome organization and retrotransposon driven molecular evolution of the endosperm *Hardness* (*Ha*) locus in *Triticum aestivum* cv Glenlea. *Molecular Genetics Genomics* 280: 467-481.
- Shewry PR, Halford NG. 2002. Cereal seed storage proteins: structures, properties and role in grain utilization. *Journal of Experimental Botany* 53: 947-958.

- Sourdille P, Perretant MR, Charmet G, Leroy P, Gautier MF, Joudrier P, Nelson JC, Sorrells ME, Bernard M. 1996. Linkage between RFLP markers and genes affecting kernel hardness in wheat. *Theoretical and Applied Genetics* 93: 580-586.
- Turner AS, Bradburne RP, Fish L, Snape JW. 2004. New quantitative trait loci influencing grain texture and protein content in bread wheat. *Journal of Cereal Science* 40: 51-60.
- Wilkinson M, Wan Y, Tosi P, Leverington M, Snape J, Mitchel RAC, Shewry PR. 2008. Identification and genetic mapping of variant forms of puroindoline b expressed in developing wheat grain. *Journal of Cereal Science* 48: 722-728.



# Evolution of GPI-Aspartyl Proteinases (Yapsines) of *Candida* spp

Berenice Parra-Ortega<sup>1,2</sup>, Lourdes Villa-Tanaca<sup>1\*</sup>  
and César Hernández-Rodríguez<sup>\*1</sup>

<sup>1</sup>*Departamento de Microbiología,  
Escuela Nacional de Ciencias Biológicas del Instituto Politécnico Nacional*  
<sup>2</sup>*Centro de Diagnóstico y Vigilancia Epidemiológica de Distrito Federal  
Instituto de Ciencia y Tecnología  
México*

## 1. Introduction

The *Candida* genus is a polyphyletic genus with at least 150 species. Nine are recognized opportunistic pathogens of humans and animals. *C. albicans* is the species most frequently isolated from human infections, followed by *Candida non-Candida* species (CNCA), as *C. glabrata*, *C. tropicalis*, *C. dubliniensis*, *C. parapsilosis*, *C. guilliermondii*, *C. lusitaniae*, *C. kefyr* and *C. krusei* (Méan et al. 2008; Pfaller & Diekema, 2007; Almirante et al. 2005; Manzano-Gayosso et al. 2000).

Some works describe the phylogenetic relationships of *Candida* genus and illustrate the limited relationship between the pathogenic *Candida* spp. The genus has been divided into: the CTG clade, which includes yeast that encodes CTG as serine instead of leucine (*C. albicans*, *C. dubliniensis*, *C. tropicalis*, *C. parapsilosis* and *C. lusitaniae*); and the WGD clade, which includes yeast that has undergone a genome duplication event (*Saccharomyces* spp., *Kluyveromyces* spp. and *C. glabrata*). Evidently, *C. glabrata* is more related to non-pathogenic yeasts, as *Saccharomyces cerevisiae*, than to the other pathogenic species (Scannell et al. 2007). *C. albicans* is a normal microorganism in humans, and colonise up to 70% of skin, mucoses, and faeces of individuals with no apparent detriment to health. However, in some circumstances, either through environmental factors or a weakening of the host immune system, a proliferation and infection by *C. albicans* arise inducing candidosis (Wei et al. 2011).

Biofilm formation, adhesion, cavitation, phenotypic switching, dimorphism, interaction with the host immune system, invasion and tissue damage are virulence factors for *C. albicans*. All these factors are related to the secreted aspartyl proteases (Sap) family, which is considered an important virulence factor and is studied as a possible target for therapeutic drug design (Naglik et al. 2004; Chaffin et al. 1998; Hube, 1998; Naglik et al. 2003, 2004, 2008).

The topic of this chapter is to understand the molecular characteristics, evolution and putative functions of glycosylphosphatidylinositol (GPI)-linked aspartyl proteases (Yps), a protein superfamily distributed among all pathogenic *Candida* species. Cell location motifs,

gene duplications, similitude, syteny, putative transcription factor binding sites and genome traits of the Yps family members are analysed by bioinformatics tools in an evolutionary context.

## 2. Aspartyl proteases

Aspartyl proteases or acid proteases (optimum activity at acidic pH) are proteins with a signal peptide in the amino-terminal site, at least one aspartic residue in the active site, and 4 cysteins (Hube & Naglik, 2001). The signal peptide is processed in the endoplasmic reticule and the protein is transported to their corresponding cell localization by the secretory pathway. The active site is formed by different amino acids. The consensus pattern described by PROSITE-EXPASY (<http://expasy.org/prosite/>) is [LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]-[GQ]-[LIVMFSTNC]-[EGK]-[LIVMFGTA], and the cysteins help the protein to the three dimensional structure by intramolecular disulfide bond (Fig. 1). According to the cell localization, aspartyl proteases could be secreted, or destined to vacuole or cell membrane by a GPI-linked site in the carboxyl-terminal residues (Alberch et al. 2006; Jones, 1991; Naglik et al. 2003).

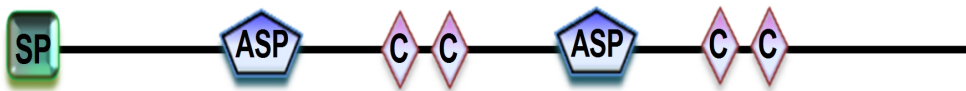


Fig. 1. Typical molecular structure of aspartyl proteases. SP: signal peptide; ASP: aspartic residue in the active site, C: cystein.

### 2.1 Secreted aspartyl proteases (Sap)

The *C. albicans* secreted aspartyl protease family comprises ten members, eight of which are proper secreted Sap1-Sap8, and two, Sap9 and Sap10, that have been reclassified as GPI-anchored aspartyl proteases (Alberch et al. 2006). Nevertheless, Sap9 and Sap10 are clearly more phylogenetically related to Sap than any GPI-anchored aspartyl proteases (Parra et al. 2009). The function of Sap in *C. albicans* has been widely studied, and these proteases are important in proteolysis to get a source of nitrogen, and are differentially regulated depending on the environmental conditions (Schaller et al. 1998; 2003; Taylor et al. 2005; Naglik et al. 2008). *SAP1-SAP3* are relevant in phenotypic switching during the opaque phase and are not expressed in the WO-1 phase (Morrow et al. 1992; White et al. 1993). Also, they are expressed when yeast colonize and damage reconstructed human epithelium, oral and vaginal, which means that these Sap are important in superficial infections (Schaller et al. 1998; 2003; Copping et al. 2005). *SAP1-SAP8* are related to tissular damage (Taylor et al. 2005). *SAP1*, *SAP3* and *SAP8* are expressed in oral and vaginal infections. On the other hand, *SAP4-SAP6* are related to systemic infections and only they are expressed in yeast and germ tube at pH 5-7 (Hube et al. 1997; Sanglard et al. 1997; White & Agabian, 1995). Meanwhile *SAP5* is important in epithelial colonization, invasion and infection (Naglik et al. 2008; Lermann & Morschhäuser, 2008).

This kind of proteases are no exclusive of *C. albicans*. Orthologous genes have been described in other closely related species, as *C. dubliniensis* (Sap1-4 and Sap7-10), *C. tropicalis* (Sapt1-12), *C. guilliermondii* (Sapg1-8), *C. parapsilosis* (Sapp1-14) and *C. lusitanae* (Sapl1-3)

(Parra et al. 2009). Particularly in *C. dubliniensis*, the expression of *SAPD3* and *SAPD4* genes is related to the infection of keratinocyte (HaCAT cells) by yeast. The number and shape of the keratinocyte cells was altered by the infection, but these effects decreased in the presence of pepstatin A, an aspartyl protease inhibitor, suggesting that the Sapd3 and 4 of *C. dubliniensis* could be considered as virulence factors, like their orthologous genes from *C. albicans* (Loaiza-Loeza et al. 2009). The function of these proteases in metabolism and pathogenesis in the rest of pathogenic species is unknown.

According to Dayhoff, protein superfamilies and families are defined as groups of related proteins that exhibit less than 50% and greater than 50% similarity, respectively. Subfamilies were defined as groups of proteins with at least 90% similarity and were often equivalent to clusters of orthologous groups (COGs) (Dayhoff, 1979). Behind this idea, the phylogeny of pathogenic *Candida* spp. Saps allows for the recognition of a superfamily with at least 12 paralogous families and nine orthologous subfamilies. In several Sap families, at least two subfamilies or orthologous groups are proposed (Parra et al. 2009).

## 2.2 Vacuolar aspartyl proteases (PrA)

The vacuole is a hydrolytic organelle similar to lysosomes in animals and is the site of non-specific degradation of cytoplasmic proteins (Robinson et al. 1988), proteins delivered via autophagy (Klionsky & Emr, 2000), or plasma membrane proteins turned over via endocytosis (Hicke, 1996). In *S. cerevisiae* the vacuole has been studied and possesses different vacuolar proteases (Table 1).

One of the most important vacuolar proteins is the proteinase A (PrA), encoded by the *PEP4* gene. Mutants in *PEP4* (*pep4*) accumulate multiple zymogens, indicating that PrA initiates processing, maturation and activation of multiple different precursors of PrB, DAP, CPY and PrA, because of their autocatalytic activity and their lack of production of dead cells in nutritional stress. Also, PrA is important in cellular response to starvation, microautophagy, proteolysis involved in cellular and vacuolar protein catabolic process, and sporulation (Palmer, 2007; Jones, 1991; Teichert et al. 1989).

The function of PrA, encoded by the *CaPEP4* gene in the metabolism of *C. albicans*, has also been studied. Null mutants of *CaPEP4* maintain their hydrolytic activity intact, clearly suggesting that *C. albicans* possesses an alternative system that compensates for the lack of this gene (Palmer, 2007). In *C. albicans*, the vacuole is important in cell differentiation, surviving into macrophages, and elimination of drugs as hygromycin B, orthovanadate and rapamycin (Palmer, 2005).

In *C. dubliniensis*, this protein could be important in carbon and nitrogen metabolism and might participate in protein degradation and precursor processing as occurs in *S. cerevisiae* (Loaiza et al. 2007). The genome-wide environmental stress response expression profile of *C. glabrata* revealed that *CgPEP4* is induced in osmotic stress and glucose starved conditions. Meanwhile, in *S. cerevisiae* no changes in the expression were observed in the same conditions (Gash et al. 2000; Roetzer et al. 2008).

Bioinformatic genomic analysis of *Candida* pathogenic species exhibited that only one version of PrA is harboured by yeast (Table 3), but apparently the *CgPEP4* gene is universally distributed among *C. glabrata* strains, as revealed by PCR multiplex in a collection of 52 *C. glabrata* clinical strains (Table 5; Fig. 3; for PCR conditions see 2.3 section). Phylogenetic analysis was performed by an alignment of PrA homologues identified *in silico* and those previously characterized. The alignment was conducted using MUSCLE in SeaView 2.4 program (Galtier et al. 1996) with default alignment parameter adjustments.

The phylogenetic analyses were performed in the MEGA4 program (Tamura et al. 2007) using Maximum Parsimony evolution. A similarity and identity matrix were computed with the MatGAT4.50.2 software (Campanella et al. 2003). The phylogenetic reconstruction and similarity of PrA reproduce the phylogenetic tree topologies of *Candida* spp. obtained with other genes, suggesting a common ancestral gene (Fig. 2; Table 2). In brief, *C. albicans* was more related to *C. dubliniensis*, followed by *C. tropicalis*, *C. parapsilosis*, *C. guilliermondii* and *C. lusitaniae*. Meanwhile, *C. glabrata* PrA was more related to *S. cerevisiae* PrA than other *Candida* species.

Name/systematic name	Gene/Protein	Access number	Function	Reference
Proteinase A YPL154C	PEP4/ PrA	NM_001183968 / NP_015171	Activities of other yeast vacuolar hydrolases	Parr et al., 2007
Carboxypeptidase Y YMR297W	CPY	NM_001182806 /NP_014026	Contributes to the proteolytic function of the vacuole	Wünschmann et al., 2007
Proteinase B YEL060C	PRB1	NM_001178875 /NP_010854	Involved in protein degradation in the vacuole and required for full protein degradation during sporulation	Teichert et al., 1989
Carboxypeptidase S YJL172W	CPS	X63068/ CAA44790	Nitrogen compound metabolic process, proteolysis involved in cellular protein catabolic processes	Bordallo & Suarez-Rendueles, 1993
Dipeptidyl aminopeptidase B* YHR028C	DAP-B	X15484/ CAA33512	Protein processing	
Aminopeptidase YKL103C	APEI	NM_001179669 /NP_012819	Catabolic processes	

Table 1. Soluble and membrane-bound \* vacuolar proteolytic system of *S. cerevisiae*.

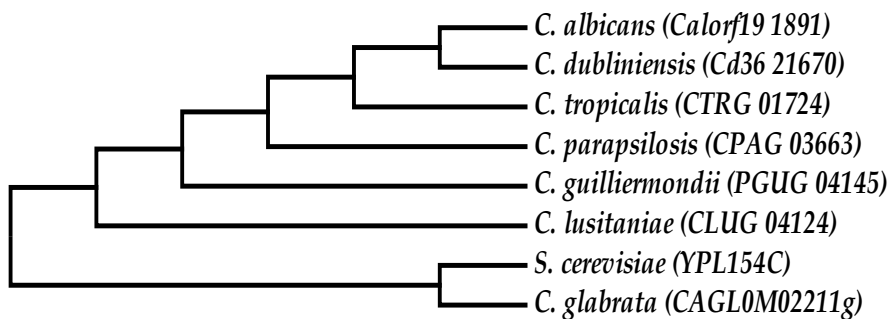


Fig. 2. Maximum Parsimony phylogenetic analysis of vacuolar aspartyl proteases (PrA) superfamily from pathogenic *Candida* spp.

	1	2	3	4	5	6	7	8
<b>1. <i>S. cerevisiae</i> YPL154C</b>		62	66	66	66	66	65	68
<b>2. <i>C. glabrata</i> CAGL0M02211g</b>	77		55	55	54	54	55	57
<b>3. <i>C. albicans</i> orf19_1891</b>	77	69		98	90	85	75	78
<b>4. <i>C. dubliniensis</i> Cd36_21670</b>	77	69	99		91	85	76	79
<b>5. <i>C. tropicalis</i> CTRG_01724</b>	76	69	97	97		86	75	77
<b>6. <i>C. parapsilosis</i> CPAG_03663</b>	76	67	91	92	93		74	78
<b>7. <i>C. guilliermondii</i> PGUG_04145</b>	79	69	84	85	85	85		78
<b>8. <i>C. lusitaniae</i> CLUG_04124</b>	78	69	86	86	87	86	87	

Table 2. Similarity and identity (UP/down) between PrA proteins from pathogenic *Candida* spp.

PrA (AN)	Amino acid residues	MM (kDa)	IP	MOTIF	Signal peptide (aa)	C
<i>C. albicans</i> Calorf19_1891	419	45.4	4.5	119-130: VILDTGSSNLWV	20	7
<i>C. dubliniensis</i> Cd36_21670	419	45.4	4.5	304-315: AAIDTGTSLLTL	14	2
<i>C. tropicalis</i> CTRG_01724	422	45.6	4.5	121-132: VILDTGSSNLWV 306-317: AAIDTGTSLLTL	24	2: 1400605-1401870-
<i>C. parapsilosis</i> CPAG_03663	428	46	4.5	127-138: VILDTGSSNLWV 312-323: AAIDTGTSLLTL	25	130: 135636-136919 -
<i>C. guilliermondii</i> PGUG_04145	409	44	4.3	109-120: VILDTGSSNLWV 294-305: AAIDTGTSLLTL	21	5: 355498-356724 -
<i>C. glabrata</i> CLUG_04124	407	43	4.3	107-118: VILDTGSSNLWV 292-303: AAIDTGTSLLTL	19	5: 46984-48204 -

Table 3. Vacuolar aspartyl proteases in pathogenic *Candida* species. (AN): Access number in the respective genome; MM: molecular mass; IP: Isoelectric Point; C: Chromosome or Contig or supercontig.

*C. glabrata* is an opportunistic haploid yeast that suffered evident and extensive reductive evolutionary events. A lot of genes involved in nitrogen metabolism, carbohydrate assimilation (saccharose, galactose, etc.), as well as sulfur, phosphor, thiamine, pyridoxine and nicotinic acid biosynthesis have been lost from the genome (Byrne & Wolfe, 2005;

Wolfe, 2006). This species produces between 15-20% of reported systemic yeast infections (Almirante et al. 2005; Manzano-Gayosso et al. 2000; Trick et al. 2002; Méan et al. 2008). *C. glabrata* is the most common yeast species isolated from patients with cancer, organ transplantation and fluconazole therapy (Safdar et al. 2001; Bodey et al. 2002). The mortality associated with *C. glabrata* in systemic infections of cancer patients is 50% and almost 100% in transplant patients (Anaissie et al. 1992; Goodman et al. 1992; Krcmery et al. 1998). This scenario is related to indiscriminate antifungal use, and to the innate resistance of *C. glabrata* (Sobel, 2006).

According to Table 4, virulence factors of *C. albicans* and *C. glabrata* are quite different. However, an evident feature is the difference in number and kind of aspartyl proteases. A total of 12 YPS genes, but no SAP genes have been detected in *C. glabrata*. Contrarily, a total of 10 SAP genes, but no YPS genes have been recognized in *C. albicans*. Clearly, the phylogenetic trees constructed with ribosomal or other gene groups include the majority of the clinical relevant *Candida* species, with exception of *C. glabrata*, which is grouped in another cluster with non-pathogenic yeasts, as *S. cerevisiae* and *Kluyveromyces* spp. This evidence suggests that the aspartyl proteases in *Candida* spp. have evolved independently as virulence factors at least two times, and possibly the amplification by duplication of SAP and YPS gene superfamilies in clinically relevant species is an example of convergent evolution.

A physiological approach could possibly contribute to the understanding of which *C. glabrata* YPS (CgYPS) genes are covering the functions of each secreted aspartyl protease of *C. albicans* under different conditions. Evidently, the comparison of virulence strategies, expression profiles, complementation of mutants, among other experiments, could suggest common and particular features and roles for all SAP and YPS genes. For now, the questions remain open. Have the function of *C. glabrata* CgYPS and SAP *C. albicans* genes functionally converged?

The transcription profile of 11 CgYPS was studied when yeast were ingested by macrophages. Apparently, CgYPS are important in survival and virulence of the yeast in macrophages, damage to mouses, Epa1 protein processing, and cell wall integrity, as occur in *S. cerevisiae*, which possesses 5 ScYPS (ScYPS1-ScYPS3, ScYPS6 and ScYPS7) (Kaur et al. 2007; Krysan et al. 2005). They are important to cell wall synthesis and glucan homeostasis, mainly ScYPS1 and ScYPS7. It seems that ScYPS3 does not have functions associated with the cell wall (Krysan et al. 2005).

*C. albicans* SAP9 and *C. glabrata* CgYPS1 genes complement the defects in the cell wall provoked by *yps1Δ* of *S. cerevisiae*. One important difference is that SAP9 complement *yps1Δ* only when SAP9 is under a heterologous and constitutive promoter from *S. cerevisiae*, while CgYPS1 complements the mutation, using its promoter (Krysan et al. 2005), evidence that supports the orthologous status proposed above for these gene pairs. As happened with ScYPS1, SAP9 gene expression increases during the stationary phase and damage of the cell wall (Monod et al. 1998; Copping et al. 2005), and protects the yeast from caspofungin (an inhibitor of  $\beta$ -1,3-glucan synthesis) (Lesage et al. 2004). Also, inhibitors of ScYps1p disable the specificity of both proteins, ScYps1p and Sap9 (Cawley et al. 2003).

Distribution of the SAP gene superfamily among *C. albicans* strains is universal (Gilfillan et al. 1998; Bautista et al. 2003; Parra et al. 2009), although one study concludes that the distribution of SAP genes in clinical strains depends on infection associated with isolation

(Kalkanci et al. 2005). Given the number of *CgYPS* in *C. glabrata* and their potential role in pathogenesis, it is important to establish the universality of *CgYPS* in *C. glabrata* populations.

Factor	<i>C. glabrata</i>	<i>C. albicans</i>	Reference
Infection sites	Oral, vaginal, bloodstream		Fidel et al. 1999
Mortality in systemic infection	urinary tract		Abi-Said et al. 1997; Krcmery, 1999
Virulence in animal models	High		Arendrup et al. 2002
Filamentation	Present		Lachke et al. 2002; Csank & Haynes, 2000
Biofilm formation	High		Castaño et al. 2006
Adherence to oral keratinocytes	Lower	Higher	Nikawa et al. 1995; Biasoli et al. 2002
Adherence to denture material	Lower	Higher	Luo & Samaranayake, 2002
Extracellular proteinase activity	Absent	Present	Chakrabarti et al. 1991
Phospholipase activity	Isolation site dependent	High	Samaranayake et al. 1994; Ghannoum, 2000
Phenotypic switching	Low	High	Brockert et al. 2003
IL-8 induction in oral keratinocytes	Pseudohyphae	True hyphae and pseudohyphae	Schaller et al. 2002
GM-CSF induction in oral keratinocytes	Weak	Strong	Schaller et al. 2002; Li et al. 2007a
Human-defensin resistance	Strong	Weak	Joly et al. 2004; Feng et al. 2005
Histatin resistance	Partially resistant	Susceptible	Helmerhorst et al. 2005
Azole resistance	High	Low	Sanglard et al. 1999
Molecules involved in adherence	20 <i>EPA</i> genes	<i>ALS</i> proteins	Castaño et al. 2005; Hoyer et al. 2001
<i>SAP</i> genes	0	10	Parra et al. 2009
<i>YPS</i> genes	12	0	Albrecht et al. 2006; Kaur et al. 2007 This work

Table 4. Comparison of virulence factors of *C. glabrata* and *C. albicans* (modified from Li, 2007b).

Our group explored the *CgYPS* gene distribution among clinical isolates (n=52) and type strains CBS138 and BG6 (N=2) by an original multiplex PCR procedure (Table 5). The yeasts were routinely grown on YPD broth and DNA was extracted using a previously reported protocol (Hoffman & Winston 1987). PCR was performed in a DNA thermal cycler 9600

(Applied Biosystems, Foster City, CA). Amplification reactions (25  $\mu$ L) were performed using a buffer containing 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 2 mM  $MgCl_2$ , 0.2 mM of each deoxynucleoside triphosphate, 0.6  $\mu$ M each primers, 4 ng/ $\mu$ L of genomic DNA, and 1.5 U/ $\mu$ L of *Taq* polymerase (Invitrogen). The PCR conditions included a denaturation step for 3 min at 94°C, followed by 38 amplification cycles consisting of 1 min at 94°C, 1 min annealing temperature and 1 min at 72°C. A final extension step was performed for 7 min at 72°C. Fig. 3 shows the amplification products of *CgYPS* gene fragments of some representative *C. glabrata* clinical strains electrophoresed in 1% agarose gels. Similar PCR conditions were used to study the universal distribution of PrA.

Gene	Primer	Location	Expected amplified fragment (bp)	T <sub>m</sub> (°C)
<b>CgYPS1</b> CAGL0M04191g	F:5'-TTCGGTGACAGTTGTATCTTGG-3' R:5'-GATAAATGAAACAAAAGACCAGCG-3'	+1326 a +1348 +1779 a +1803	477	55
<b>CgYPS2</b> CAGL0E01419g	F:5'-ACTCAACTGTITTTAACTTCGGTGGTGC-3' R:5'-TAGCATGGAGAGTAGGATGTTAAACACC-3'	+1234 a +1262 +1743 a +1770	536	61
<b>CgYPS3</b> CAGL0E01727g	F:5'-AAAGCAAGTCGTCGATGTCATCG-3' R:5'-TTGCAACTAACTAAAGTGGTGC-3'	+951 a +973 +1580 a +1603	652	58
<b>CgYPS4</b> CAGL0E01749g	F:5'-TTCGTGTTACCAGCAAAGGTTGC-3' R:5'-TTAATGTAGTTCTCTTACGGAGAGC-3'	+933 a +956 +1411 a +1435	502	55
<b>CgYPS5</b> CAGL0E01771g	F:5'-TATACATATATGCCAAGCAGCGTTGC-3' R:5'-AACAAGCGAGTAACTGCTGATAAAGC-3'	+934 a +959 +1528 a +1553	619	58
<b>CgYPS6</b> CAGL0E01793g	F:5'-ACCAGAAGGTAGCTGCATTAATCG-3' R:5'-AATGGTAGCTAATATGGCAGCAACG-3'	+887 a +910 +1542 a +1518	631	58
<b>CgYPS7</b> CAGL0A02431g	F:5'-TATGGGACCAATCTATATAACGTCC-3' R:5'-TAAGTAGCATACGGTATGTAGCCC-3'	+831 a +855 +1404 a +1427	596	55
<b>CgYPS8</b> CAGL0E01815g	F:5'-TTGGGATTACAGGGTAATGATGC-3' R:5'-AACTCTTTTTGAAGGTCAAAACGCG-3'	+856 a +878 +1457 a +1482	626	58
<b>CgYPS9</b> CAGL0E01837g	F:5'-TTCCGTAATGTGACTGATTCATGG-3' R:5'-ATCATAATGAGTATGGCAGAGTTGGC-3'	+1071 a +1096 +1510 a +1535	464	58
<b>CgYPS10</b> CAGL0E01859g	F:5'-TAATAAGACGGAAGCCATCAGACTGC-3' R:5'-TTGTAATIGCTGCTAGTACTAGGACG-3'	+978 a +1003 +1479 a +1504	526	58
<b>CgYPS11</b> CAGL0E01881g	F:5'-TTGGTGTCCCATACAAGGAAATGGTC-3' R:5'-AATCCACAAG ACCAGCAACA GGATAGC-3'	+1100 a +1125 +1495 a +1521	421	61
<b>CgYPS12</b> CAGL0J02288g	F:5'-AATTGCACATGAAGATTCCGTGCG-3' R:5'-TATCAGTTATTGTAGCAGTTACTGGC-3'	+1001 a +1025 +1542 a +1567	566	58
<b>CgPEP4</b> CAGL0M02211g	F:5'-TATCTGAAGAGTGTCAATGACCCAGC-3' R:5'-TACAGCCTCAGCTAAACTGACAACATTGG-3'	+691 a +716 +1208 a +1236	545	58

Table 5. Primer pairs used for conventional multiplex PCR of *C. glabrata* YPS genes. Bp, Base pair; T<sub>m</sub>, Melting temperature.



The universality of the 12 *CgYPS* genes among all *C. glabrata* clinical isolates and type strains was confirmed (Fig. 3), which suggests that all *CgYPS* are important to yeast life cycle as pathogen or commensal, and probably are differentially regulated according to each environmental condition, as occurs with *C. albicans* *SAP*.

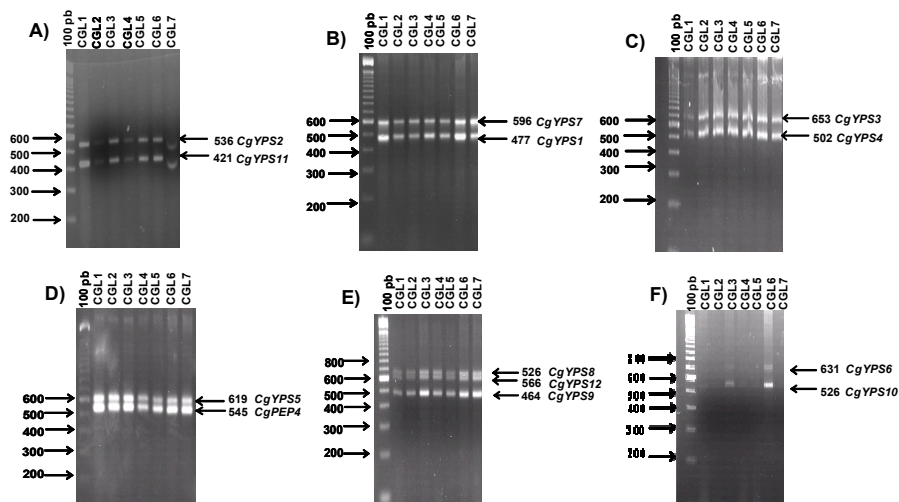


Fig. 3. Amplification of *CgYPS* *C. glabrata* gene fragments by multiplex PCR. A, *CgYPS2* and *CgYPS11*; B, *CgYPS7* and *CgYPS1*; C, *CgYPS3* and *CgYPS4*; D, *CgYPS3* and *CgPEP4*; E, *CgYPS8*, *CgYPS12* and *CgYPS9*; F, *CgYPS6* and *CgPEP10*.

### 2.3 *YPS* genes in clinically relevant *Candida* species

The genome sequence projects of *Candida* species allows for the exploration of whether *YPS* genes are harboured in these opportunistic pathogen yeasts. *C. dubliniensis* sequences were obtained from the Sanger Institute Microorganisms Sequencing Group (<http://www.sanger.ac.uk/sequencing/Candida/dubliniensis/>). Sequences from *C. guilliermondii*, *C. lusitaniae*, *C. tropicalis* and *C. parapsilosis* were obtained from ([http://www.broad.mit.edu/annotation/genome/candida\\_group/MultiHome.html](http://www.broad.mit.edu/annotation/genome/candida_group/MultiHome.html)). The GenBank database (<http://www.ncbi.nlm.nih.gov>) was also used. The detection was made by using the previous *YPS* and *SAP* genes detected in *S. cerevisiae* (<http://www.yeastgenome.org>), *C. glabrata* (<http://cbi.labri.fr/Genolevures/elt/CAGL>) and *C. albicans* (<http://www.candidagenome.org>) genomes, and the proteins detected by BLAST analysis in NCBI. Also, the different patterns of motif that could be obtained were used as a new query. In *C. lusitaniae* and *C. guilliermondii* only one *YPS* was detected. Meanwhile in *C. dubliniensis* and *C. albicans* four *YPS* genes were detected, in *C. tropicalis* two, and in *C. parapsilosis* six. Theoretical isoelectric point, molecular weight and amino acid content were calculated using Antheprot 2000 version 5.2 (Table 6).

Prediction of motif sequences was performed with PROSITE (<http://www.expasy.org>) (Falquet et al. 2002). Some of the proteins possess a typical molecular structure of aspartyl proteases, but others have some differences in composition (Fig. 4; Table 6). Some of them possess high Ser/Thr content in the amino terminal, suggesting that this zone is exposed at the surface of the protein. The presence of Ser/Thr in the carboxyl terminal in almost all *YPS*

is postulated to be heavily O-glycosylated. The exact function of this Ser/Thr-rich domain in yapsins has not been investigated. However, O-mannosylation is important for proper cell-wall biogenesis and integrity. It has also been proposed that clustered O-glycans create rigid stalks that keep protein domains away from membranes or wall surfaces (Lipke & Ovalle, 1998).

Yps (AN)	Amino acid residues	MM (kDa)	IP	MOTIF	Signal peptide (aa)	C
ScYps1 YLR120C	569	60	4.5	98-109: VLVDTGSSDLWI 368-379: ALLDSGTTLTLYL	21	XVII
ScYps2 YDR144C	596	64.2	4.3	96-107: VLVDTGSSDLWV 356-368: VLLDSGTTISYM 496-570: SER	18	IV
ScYps3 YLR121C	508	54.5	8.4	78-89: VLLDTGSADLWV 285-296: ALLDSGTTLTLYL 439-470: THR	20	XVII
ScYps6 YLR039C	537	58.2	3.9	82-93: LQLDTGSSDMIV 321-32: VMLDSGTTFSYL	24	IX
ScYps7 YDR349C	596	64.4	4.6	71-82: LLVDVIIQPYINL 318-329: ALLDSTSSVSYL	16	IV
ScBar1 YIL015W	587			60-71: VLFDTGSADEFWV 284-295: VLLDSGTSLNA		
CgYps1 CAGL0M04191g	601	63.8	5.0	88-99: VLVDTGSSDLWI 375-386: ALLDSGTTLTLYL	18	M
CgYps2 CAGL0E01419g	591	63.2	4.4	82-92: LLLDTGSSDMWV 366-377: ALLDSGTTVSYL	18	E
CgYps3 CAGL0E01727g	539	58.9	6.4	66-79: VQLDTGSSDLWF 305-316: VLLDTGTTLAYA	14	E
CgYps4 CAGL0E01749g	482	53.2	8.4	65-76: VQLDTGSSDLWF 303-14: TLLDTGVTTSVL	15	E

Yps (AN)	Amino acid residues	MM (kDa)	IP	MOTIF	Signal peptide (aa)	C
CgYps5 CAGL0E01771g	519	57.2	5.5	66-77: VQLDTGSSDLWF 304-315: ALLDTGTTYTYM 60-70: LECT	15	E
CgYps6 CAGL0E01793g	516	55.9	4.6	65-76: VQLDTGSADLWF 301-312: ALIDSGTTISEF 62-68: LECT	15	E
CgYps7 CAGL0A02431g	587	63.4	4.7	64-75: LGLGLAQPYVWV 302-313: VLLDPSFALSYL	18	A
CgYps8 CAGL0E01815g	519	56.7	6.8	65-76: VQLDTGSSDLWF 304-315: ALLDSGTTLTVV	15	E
CgYps9 CAGL0E01837g	521	56.9	5.1	65-76: LQIDTGSSDLFV 300-311: TLLDSGSTISLL	16	E
CgYps10 CAGL0E01859g	505	55.3	7.3	61-72: AQLDTGSSDLWF 298-309: ALFDSGTSYSYV	13	E
CgYps11 CAGL0E01881g	508	55.6	5.0	63-74: LLVDTGSSDFWV 310-321: ALLDTGSTDTHL	29	E
CgYps12 CAGL0J02288g	541	59.5	4.6	68-79: LVLDTGSSDLWV 279-290: ALLDTGSTLIEL 448-495: SER	19	J
orf19_852	365	39.6	5.4	73-84: LAADTGSWLIQI 245-256: YTIDTGGRYGFL	17	2
orf19.6481	702	75.9	4.4	159-170: LRLDLIQPEIWV 406-417: VLLDSRASNFYL 565-662: SER	20	7

Yps (AN)	Amino acid residues	MM (kDa)	IP	MOTIF	Signal peptide (aa)	C
orf19_853	364	39.1	5.7	72-83: LSIDTGSWLTHI 244-255: YTLDTGGGTGFL 42-44: RGD	17	2
orf19_2082*	436	47.7	3.8	53-64: VIVDSGSSDLMI 229-240: YQIDSGTNGFVP	14	2
Cd36_18360				72-82: VVII-1DTGSWLTHI 848-859: YTLDTGGGNGYL	17	2
Cd36_18370	365	40	5.6	73-84: IAADTGSWLTQI 246-257: YTMDTGGGYGYL	17	2
Cd36_72090	697	76.6	4.6	149-150: LRDLIQPEIWVM 402-412: VILDSRASNFY	13	7
Cd36_15430	442	48.7	4.2	60-71: VII-1VDSGSSDLMI 236-47: YQIDSGTNGFVP	27	2
CTRG_05014	690	74.8	4	151-162: LRDLIQPEIWV 401-412: VLIDSRSSYFYL	20	7: 407395- 409464 -
CTRG_01112	432	47.9	3.8	55-66: VII-1VDSGSSDLMI 232-243: YQIDSGSNGFLP 392-423: THR	20	2: 57814- 59109 -
CPAG_04785	369	40.5	4.5	72-83: VMIDTGSWRLNV 245-256: IGIDSGNPRLAF	20	139:296423 -297529 -
CPAG_04801	374	40.5	5.7	251-262: LALDTGNPGIGL 76-77: VFIDTGSWALNF	19	139:334234 -335355 -
CPAG_04802	371	40.4	6.1	76-87: VVII-1DTGSWALNF 248-259:	19	139:337822 -338934 +

Yps (AN)	Amino acid residues	MM (kDa)	IP	MOTIF	Signal peptide (aa)	C
				LAFDTGSAGLIL		
CPAG_03253	366	40.9	6.4	74-85: VLLDTASTVLNV 246-257: VLHDSGTPTMEL	15	126: 70630-71727 +
CPAG_02564	366	40.5	6.5	74-85: VLLDTASIVLVN 246-255: VLHDSGTPTMAL	15	116:138449-139546 -
CPAG_04713	700	75.5	4.5	157-168: LRLDLIQPEVWV 417-428: VLLDSRILYSYL 19-55: SER	27	139:123872-125971 +
PGUG_04882	583	63.5	4.1	81-92: LRLDLTQPEIWW 224-235: LVQQGVIIKSSAY	16	6: 496161-497909
CLUG_00903	722	74	3	63-74: VLLDTGSSDLWV 275-286: ALLDSGTSLQYL 470-701: SER 540-638: THR	14	1: 1836367-1838532 +

Table 6. Aspartyl proteases GPI- linked to cell membrane in pathogenic *Candida* spp. AN: Access number in the respective genome; MM: molecular mass in kilodaltons (kDa); IP: Isoelectric Point; C: Chromosome/Contig or supercontig; the atypical amino acids in the PROSITE motif are shown in black (Eukaryotic and viral aspartyl protease active site).

The presence of a GPI attachment site, a characteristic feature of the yapsin family, was determined with big-PI predictor ([http://mendel.imp.univie.ac.at/gpi/gpi\\_server.html](http://mendel.imp.univie.ac.at/gpi/gpi_server.html)), and GPI-SOM. GPI-anchor signals were identified by a Kohonen Self Organizing Map (<http://gpi.unibe.ch/>). A total of 36 protein sequences were analyzed, but GPI sites were recognized only in 21 proteins. GPI sites were not detected in ScYps2 and CgYps2, although both proteins have been previously confirmed as Yps proteins. The software programs must be enhanced, but an experimental approach to confirm the cell location is necessary. PSORTII (<http://www.psорт.org/>) and Softberry (<http://www.softberry.com>) programs were used to predict subcellular localization. All proteins detected seem to be extracellular, which could be because of the presence of a signal peptide in the amino terminal extreme. Nevertheless during their synthesis, yapsins are cotranslocated and modified by the addition of GPI to the lumen of the endoplasmic reticulum (ER). Then proteins are glycosylated in Golgi apparatus, associated to membrane vesicles and sent to plasma membrane or the cell wall (Mayor & Rieaman, 2004; Caro et al. 1997). Softberry program was also used to find exons, which were absent in all genes studied. A search was made for

internal protein sequence repeats to detect possible internal duplication events, but none were detected by TRUST (Szkarczyk & Heringa, 2004) even though it is likely were not

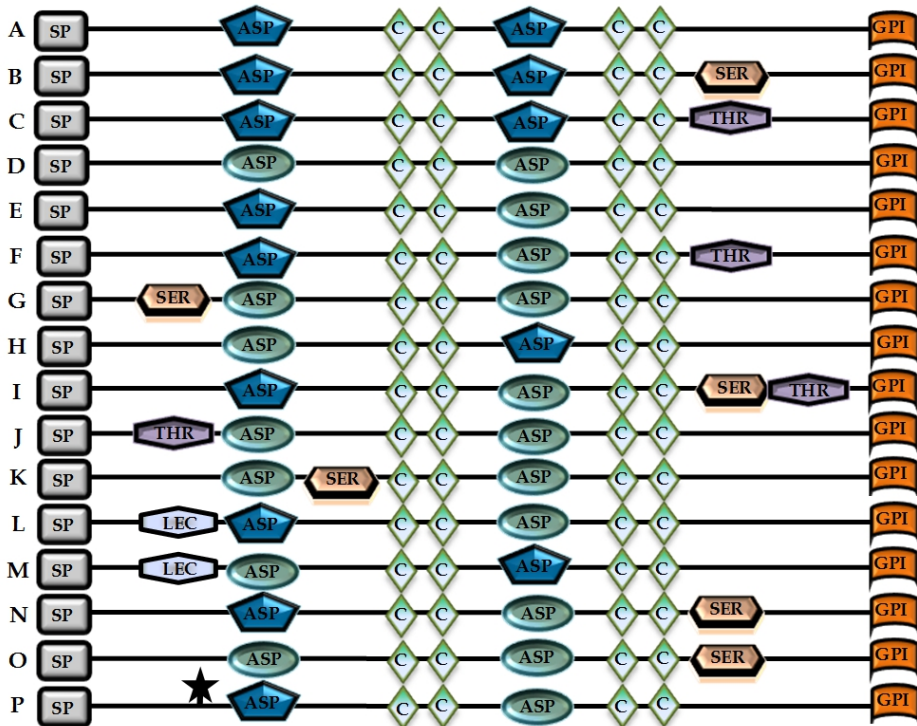


Fig. 4. Motifs of *Candida* spp. GPI-anchored aspartyl proteases (Yps). Rectangle boxes (SP): amine terminal signal peptide; pentagon (ASP): aspartyl protease domains in agreement with PROSITE; circles (ASP): atypical aspartyl protease domains proposed as [LIVMFGACTPSYF]-(LIVMTADNQSFH)-(LIVFSAE)-D-(STP)-(GS)-(STAV)-(STAPDENQY)-X-(LIVMFSTNCGQ)-(LIVMFGTAW); hexagons: serine (SER), threonine (THR), lecithin (LEC) rich regions; star: RGD motif; rhombus (C): cysteine residues, semicircles. Ca, *C. albicans*; Cd, *C. dubliniensis*; Cg, *C. glabrata*; Cgu, *C. guilliermondii*; Cl, *C. lusitanae*; Cp, *C. parapsilosis*; Ct, *C. tropicalis*; Sc, *S. cerevisiae*. **A)** ScYps1 (YLR120C), ScYps6 (YLR139C), CgYps1 (CAGL0M04191g), CgYps2 (CAGL0E01419g), CgYps11 (CAGL0E01881g); **B)** ScYps2 (YDR144C); **C)** ScYps3 (YLR121C); **D)** ScYps7 (YDR349C), CdYps (Cd36\_18370), CpYps (CPAG\_04785), CpYps (CPAG\_04801), CpYps (CPAG\_04802), CpYps (CPAG\_03253), CpYps (CPAG\_02564), CguYps (PGUG\_04882), CgYps3 (CAGL0E01727g), CgYps4 (CAGL0E01749g), CgYps7 (CAGL0A02431g), CgYps9 (CAGL0E01837g), CaYps (orf19\_852), CdYps (Cd36\_72090); **E)** CdYps (Cd36\_15430), CaYps (orf19.2082); **F)** CtYps (CTRG\_01112); **G)** CpYps (CPAG\_04713); **H)** CgYps8 (CAGL0E01815g), CgYps10 (CAGL0E01859g); **I)** ClYps (CLUG\_00903); **J)** CtYps (CTRG\_05014); **K)** CgYps5 (CAGL0E01771g), CgYps6 (CAGL0E01793g); **L)** CgYps12 (CAGL0J02288g); **M)** CaYps (orf19.6481); **N)** CaYps (orf19\_853), CdYps (Cd36\_18360).

detected by TRUST (Szklarczyk & Heringa, 2004) even when it is likely that the Yps and Sap superfamilies have duplicated aspartyl protease motifs.

The analysis of possible evolutive and molecular events that has given place to the presence of different numbers of YPS in each pathogenic *Candida* species was made to establish the COGs between Yps. Phylogenetic analysis was performed by an alignment of YPS homologues identified *in silico* and those of the previously characterized. The alignment was carried out using MUSCLE in SeaView 2.4 program (Galtier et al. 1996) with default alignment parameter adjustments. The phylogenetic analyses were performed in the MEGA4 program (Tamura et al. 2007) using minimum evolution computed with the Poisson correction. A similitude and identity matrix were computed with the MatGAT4.50.2 software (Campanella et al. 2003). To corroborate support for the branches on trees, bootstrap analysis (1,000 replicates) was performed. Synteny analysis was made to recognize the putative COGs (Fig. 5).

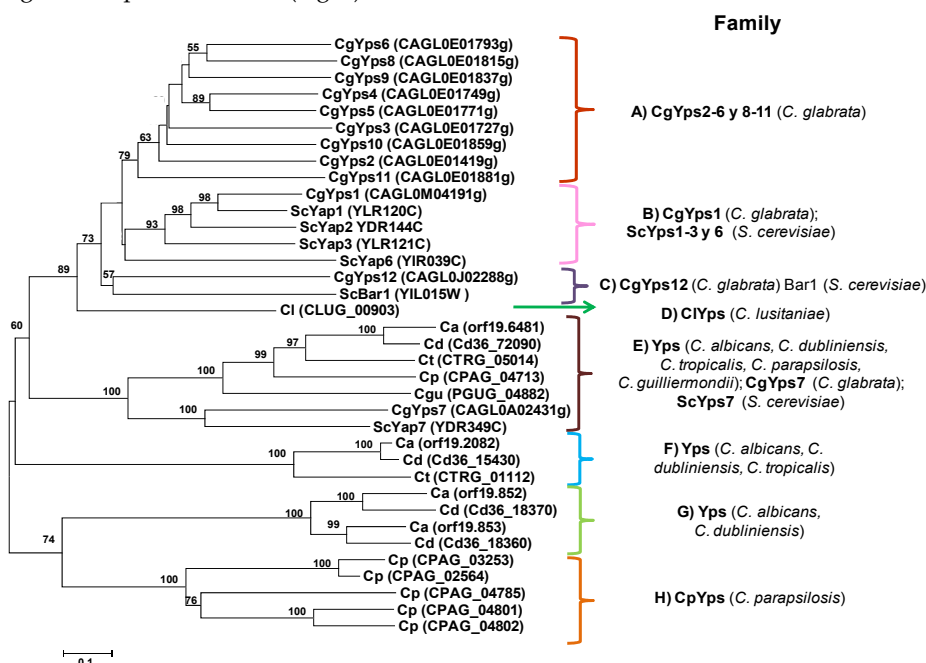


Fig. 5. Minimum evolution phylogenetic tree of GPI-anchored aspartyl proteinase (Yps) superfamily of opportunistic pathogenic *Candida* species. Ca, *C. albicans*; Cd, *C. dubliniensis*; Cg, *C. glabrata*; Cgu, *C. guilliermondii*; Cl, *C. lusitaniae*; Cp, *C. parapsilosis*; Ct, *C. tropicalis*; Sc, *S. cerevisiae*. Bootstrap values > 50% are on branches. Curly brackets and arrows indicate the Yps protein families defined by phylogenetic relationships, similitude percentage (> 50%), synteny and motif array. Yps are grouped into 8 families. Family A, CgYps2-6 and 8-11; family B, CgYps1, ScYps1-3 and ScYps6; family C, CgYps12 and ScBar1; family D, ClYps (*C. lusitaniae*); family E, CgYps7, ScYps7, CaYps (orf19.6481), CdYps (Cd36\_72090), CtYps (CTRG\_05014), CpYps (CPAG\_04713) and CguYps (PGUG\_04882); family F, CaYps (orf19.2082), CdYps (Cd36\_15430) and CtYps (CTRG\_01112); family G, CaYps (orf19.852), CdYps (Cd36\_18370), CaYps (orf19.853) and CdYps (Cd36\_18360); family H, CpYps (CPAG\_03253, CPAG\_02564, CPAG\_04785, CPAG\_04801 and CPAG\_04802).

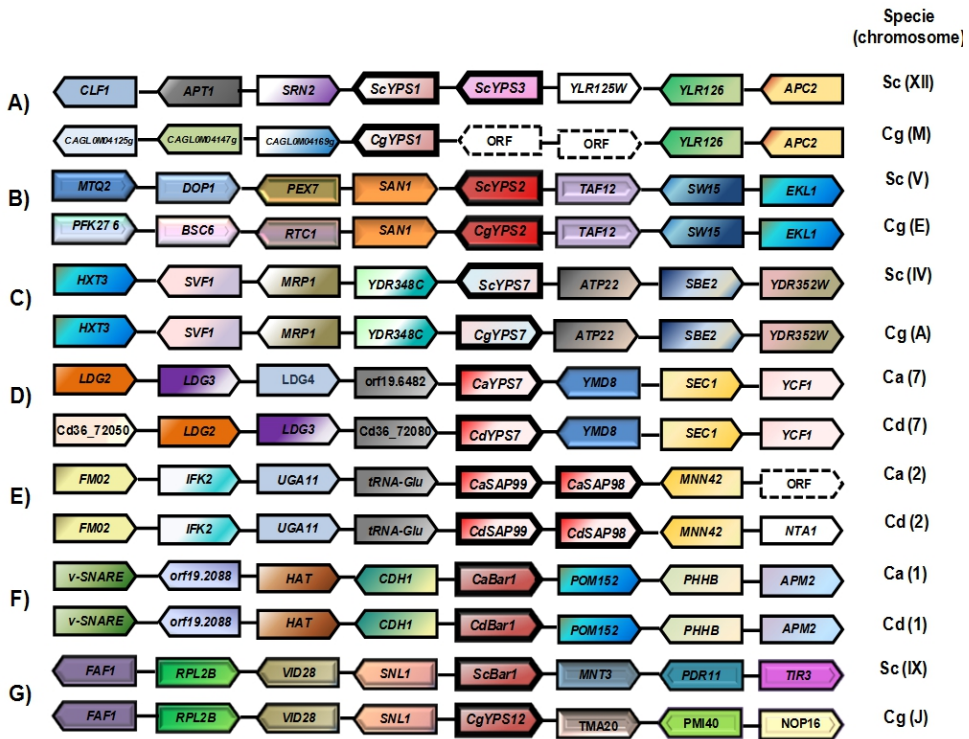


Fig. 6. Synteny of YPS genes of *S. cerevisiae* (Sc), *C. glabrata* (Cg), *C. albicans* (Ca) and *C. dubliniensis* (Cd). **A)** *ScYPS1* and *CgYPS1*; **B)** *ScYPS2* and *CgYPS2*; **C)** *ScYPS7* and *CgYPS7*; **D)** *CaYPS7* (orf19.6481) and *CdYPS7* (Cd36\_72090); **E)** *CaYPS* (Sap99), *Cd* (orf19.853 and Sap98, orf19.852); **F)** *CaYPS* and *CdYPS* (Bar1); **G)** *CgYPS* and *ScBar1*. *CgYPS1*, *CgYPS7*, *ScYPS1*, *ScYPS3*, *CaYPS7*, *CaSAP98*, *CaSAP99* and *BAR1* are GPI anchored aspartyl proteases; *APC2* and *CAGL0M04235g*, subunit of the anaphase-promoting; *APT1*, acyl-protein thioesterase; *ATP22*, mitochondrial inner membrane protein; *CAGL0M04147g*, similar to low affinity vacuolar membrane, is a localized monovalent cation/H<sup>+</sup> antiporter protein; *CAGL0M04169g*, similar to cell wall glycoprotein involved in beta-glucan assembly; *CDH1*, cell-cycle regulated activator of the anaphase-promoting complex/cyclosome (APC/C); *CLF1*, crooked neck-like factor; *DOP1*, protein essential for viability; *EKL1*, ethanolamine kinase; *FAF1*, protein required for pre-rRNA processing and 40S ribosomal subunit assembly; *HAT*, histone acetyltransferase; *HXT3*, low affinity glucose transporter of the major facilitator superfamily; *LDG3* and *LDG4*, leucine, aspartic acid, glycine rich; *MNN42*, putative positive regulator of mannosylphosphate transferase; *MNT3*, alpha-1,3-mannosyltransferase; *MTQ2*, S-adenosylmethionine-dependent methyltransferase; *MRP1*, mitochondrial ribosomal protein of the small subunit; *NOP16*, constituent of 66S pre-ribosomal particles; *NTA1*, amidase; orf19.2088, shared subunit of DNA polymerase (II) epsilon and of ISW2/yCHRAC chromatin accessibility complex; *PDR11*, ATP-binding cassette transporter, *PEX7*, peroxisomal signal receptor; *PFK27*, 6-phosphofructo-2-kinase; *PHHB*, transposon mutation affects filamentous growth; *PMI40*, mannose-6-phosphate



isomerase; *POM152*, nuclear pore membrane glycoprotein; *RPL2B*, protein component of the large ribosomal subunit; *SAN1*, ubiquitin-protein-ligase; *SBE2*, protein involved in the transport of cell wall components from the Golgi to the cell surface; *SEC1*, Sm-like protein involved in docking and fusion of exocytic vesicles through binding to assembled SNARE complexes at the membrane; *SNL1*, putative protein involved in nuclear pore complex biogenesis and maintenance; *SRN2*, component of the ESCRT-I complex; *SVF1*, protein with a potential role in cell survival pathways; *SWI5*, transcription factor that activates transcription of genes expressed at the M/G1 phase boundary and in G1 phase; *TAF12*, subunit (61/68 kDa) of TFIID and SAGA complexes; *TIR3*, cell wall mannoprotein of the Srp1p/Tip1p family of serine-alanine-rich proteins; *TMA20*, protein associated with ribosomes with a putative RNA binding domain; *UGA11*: gamma-aminobutyrate transaminase (4-aminobutyrate aminotransferase); *VID28*, protein involved in proteasome-dependent catabolite degradation of fructose-1,6-bisphosphatase (FBPase); *v-SNARE*, component of the vacuolar SNARE complex involved in vesicle fusion; *YCF1*, putative glutathione S-conjugate transporter; *YLR126C*, protein with similarity to glutamine amidotransferase proteins; *YMD8*, putative nucleotide sugar transporter; ORF, *APM2*, *BSC6*, *CAGL0M04125g*, *Cd36\_72050*, *Cd36\_72080*, *FM02*, *IFK2*, *orf19.6482*, *RTC1*, tRNA-Glu, *YDR352W*, *YDR348C* and *YLR125W* and ORF, unknown predicted open reading frame.

The lack of *SAP* genes and the expansion of 12 *CgYPS* genes in *C. glabrata*, and the extended family of *SAP* genes in *C. albicans* support the hypothesis that both protein superfamilies are an example of convergent evolution. Although more research is necessary to reach definite conclusions, apparently *YPS* of *C. glabrata* and *SAP* of *C. albicans* have developed some equivalent physiological functions and roles in virulence. The rest of pathogenic *Candida* species are less virulent, and, curiously, harbour less genes in their genomes than *C. albicans*. These facts lead to the supposition that *SAP* and *YPS* have evolved in an independent way for at least 700 million years. However, more *SAP* duplication events have happened in *C. albicans* (Parra et al. 2009).

Phylogenetic analyses of Yps deduced protein sequences of *Candida* spp. and *S. cerevisiae* allow for the definition of 8 Yps families, A-H (Fig. 5). In particular, *CgYps1-12* proteins of *C. glabrata* were clustered in four families. Family A was constituted exclusively of nine Yps of *C. glabrata* (*CgYps2-6* and *CgYps8-11*) encoded in chromosome E. With exception of *CgYps2*, all codifying genes of these proteins are organized in tandem, and possibly derived from at least eight recent duplication events that occurred exclusively in the *C. glabrata* genome. Apparently these recent duplications led to the emergence of a paralogous gene family with novel or slightly different functions. No pseudogenes were detected in *CgYPS1-11* genes, but in their deduced proteins a moderate amino acid similitude (48-53%) and identity (36-38%) were retained. Frequently, very high similitudes are maintained by concerted evolution in paralogous members of some multigene families (László, 1999). However, in *CgYPS* genes, this evolutive phenomenon is not evident. Previously, *CgYPS4* and *CgYPS11* were recognized as GPI anchored aspartyl proteases (Kaur et al., 2007), but comparative studies of the regulatory region and expression of each *CgYPS* genes are necessary to clearly define the physiological role and orthology relationships of each gene. Family B was formed by a set of Yps proteins, detected exclusively in *S. cerevisiae* (*ScYps2-3* and *ScYps6*), and a highly similar putative orthologous pair (*ScYPS1/CgYPS1*) (Fig. 6A). Also, the partial synteny observed between the *ScYPS2/CgYPS2* gene pair supports the hypothesis that those protein-coding genes are probable orthologous (Fig. 6B). Family C was

integrated by CgYps12 and ScBar1 of *S. cerevisiae*, a putative orthologous pair with low similitude synteny, but with a clear ancestor-descendant relationship (Fig. 6G). Finally, family E was formed by a representative of each *Candida* spp. Yps, CgYps7 and ScYps7. This family forms a sub tree with the same topology as those phylogenies constructed with ribosomal and other protein sequences (Diezman *et al.*, 2004). The CgYPS7 and ScYPS7 genes exhibited an extensive synteny (Fig. 6C), but no synteny with CaYPS (orf19.6481) and CdYPS (Cd36\_72090) was observed (Fig. 6D). In *C. albicans* and *C. dubliniensis* genome databases these YPS are described as ScYPS7 orthologous genes (Schaefer *et al.*, 2007). Nevertheless, both YPS exhibited low similarity with ScYPS7 (37.2-38.7%) and no-synteny. The final decision to consider family E as an orthologous family will depend on comparative analyses of functional features not yet performed.

Families C, F, G and H have not any *C. glabrata* or *S. cerevisiae* Yps representative protein. Families C and H were formed only by one CIYps gene of *C. lusitaniae* and seven CpYps genes of *C. parapsilosis*, respectively (Fig. 5). Curiously, *C. lusitaniae* is the species that harbours the fewest CIYPS (n=1) and SAP (n=3) genes, and its isolation frequency from clinical samples, as well as its virulence, are lower than the other *Candida* species (Abi-Said *et al.* 1997). This evidence supports a hypothesis of relevance of aspartyl proteases in virulence. That is, species with numerous aspartyl proteases in virulence; species with broad aspartyl proteases are more virulent than those with a limited number of these proteins.

Family F harboured *C. albicans*, *C. dubliniensis* and *C. tropicalis* yapsins organized congruently according to the ribosomal phylogenetic tree. The *C. albicans* CaBar1 (orf19.2082) and *C. dubliniensis* CdBar1 (Cd36\_15430) gene, found in family F, has been described as orthologous to *S. cerevisiae* BAR1 (Schaefer *et al.*, 2007) found in family C. In both species, *C. albicans* and *S. cerevisiae*, the protein is involved in alpha pheromone degradation and secreted to the periplasmic space of mating alpha-type cells. These proteins help cells find mating partners by cleaving and inactivating the alpha factor, which allows cells to recover from alpha-factor-induced cell cycle arrest (Mackay *et al.*, 1988). The *in silico* analysis performed in this work established that these proteins and the Bar1 from *C. dubliniensis* are extracellular, but anchored to the cell wall or cell membrane. Also, phylogenetic analysis shows that Bar1 from *C. albicans* and *C. dubliniensis* belongs to the Yps superfamily, with a similarity of 40%, and are not grouped with CgYps12 of *C. glabrata* (CgYps12 or CgBar1) and Bar1 of *S. cerevisiae*. The reason for which an aspartyl protease, that apparently is secreted, is grouped with the yapsins superfamily could be a mistake in the cell location method because almost all software use the signal peptide, transmembranal regions, and the GPI site in the C-terminal, to predict the cell location. In *C. albicans* it has been detected that aspartyl proteases are associated with the plasmatic membrane, or to both the plasmatic membrane and cell wall. This makes the experimental corroboration of the cell location necessary. The Bar1 protein of *C. albicans* has been described as a protein with three domains: 2 aspartyl protease domains and another unidentified. Apparently, this GPI-membrane anchored domain determines that Bar proteins are not secreted, but anchored to cellular membranes, and their two active sites are oriented to cellular membranes, and their two active sites are oriented to the exterior to inactivate alpha pheromone, which is secreted by Mat-alpha cells. In *C. albicans*, the degradation of secreted alpha pheromone is not exclusive to Bar1. CaYPS7 (orf19.6481) of family E also encodes for this function with lesser efficiency (Schaefer *et al.*, 2007). This physiological redundancy has not been demonstrated in *S. cerevisiae* ScYps7. *C. albicans* can mate under some *in vitro* and *in*

*vivo* conditions when alpha pheromone is degraded (Hull *et al.*, 2000; Magee & Magee, 2000) and *C. glabrata* harbours homologous genes of *S. cerevisiae* that control the mating (Srikantha *et al.*, 2003). Nevertheless, in *C. glabrata* a cell cycle has not been demonstrated, and the participation of CgYps7 of *C. glabrata* in alpha pheromone inactivation has not been demonstrated. No possible gene orthologous to possible gene orthologous to ScBar1 was detected in *C. guilliermondii*, *C. lusitaniae*, *C. parapsilosis*, *C. tropicalis*, *C. guilliermondii* or *C. lusitaniae*. All these yeasts have a heterothallic sex cycle (cross-mating only), but *C. parapsilosis* and *C. tropicalis* mating has never been observed (Butler *et al.*, 2009).

Family G is formed by two *C. albicans*/*C. dubliniensis* Yps protein pairs with high similitude (>88%), located in tandem in chromosome 2 and with very similar synteny. All this data is evidence from the recent speciation of both species (Fig. 6E). According to the *Candida* genome database (<http://www.candidagenome.org/cgi-bin/locus.pl?locus=orf19.852>) Cal orf19.852 and Cdu Cd36\_18370 sequences are described as CaSAP98 and CdSAP98 genes, respectively, and have their best hits with PEP4 of *S. cerevisiae* (Pra protein). *S. cerevisiae* PrA is a vacuolar protease, and clearly *C. albicans*/*C. dubliniensis* Yps are not phylogenetically grouped with PrA. In our opinion no orthology relationship among these proteins exists. Cal orf19.853 and Cdu Cd36\_18360 formed a second pair, described as CaSAP99 and CdSAP99 genes, which had their best hits with ScYPS3 of *S. Cerevisiae*. Similarly, it is clear that CaSAP99 has no synteny, phylogenetic relationship, or possible common physiological role with ScYPS3.

### 3. Conclusion

Why have *C. albicans*/*C. dubliniensis* and *C. glabrata*/*S. cerevisiae* been suffering some genetic duplication events in their Sap and Yps superfamilies? This is something that has not been resolved, but it is clear that the decrease in virulence in null mutants, in both CaSAP and CgYPS, endorse the idea that the presence and expansion of SAP and YPS families is necessary for adaptation to the host, and therefore for survival and virulence. Also, species with broad aspartyl protease families are more virulent than those with a limited number of these proteins. *C. glabrata* belongs to a phylogenetic group with no pathogenic yeast, and its virulence attributes could be evolving independently from the CTG clade, where *C. albicans* is the main opportunistic pathogenic species. The expansion of the CgYPS gene superfamily of *C. glabrata* maintains a parallelism with the expansion of the SAP gene superfamily of *C. albicans*, and constitutes a possible example of convergent evolution. The transition from a commensally life style to a successful opportunistic pathogen could be related to gene expansion that encodes for each kind of aspartyl protease. A lot of experimental methodologies must be performed to recognize the orthologous gene families, as well as the virulence, participation and transition commensal-pathogen roles of aspartyl proteases, including Sap and Yps.

### 4. Acknowledgment

We are grateful for the financial support from CONACyT-CB-13695, CONACyT-69984, SIP201005214 and SIP20113066. BPO is a fellow of CONACyT and PIFI-IPN. Thanks to Dr. Bernard Dujon (Institut Pasteur and Université Pierre et Marie Curie) for donating strains. Thanks also to Bruce Allan Larsen for reviewing the use of English.

## 5. References

- Abi-Said, D., Anaissie, E., Uzun, O., Raad, I., Pinzcowski, H. & Vartivarian, S. (1997). The epidemiology of hematogenous candidiasis caused by different *Candida* species. *Clin Infect Dis*. Vol. (24): 1122-1128.
- Albrecht, A., Felk, A., Pichova, I., Naglik, J., Schaller, M., de Groot, P., MacCallum, D., Odds, F., Schafer, W., Klis, F., Monod M. & Hube, B. (2006). Glycosylphosphatidylinositolanchored proteases of *Candida albicans* target proteins necessary for both cellular processes and host pathogen interactions. *J Biol Chem*. Vol. (281): 688-694.
- Almirante, B., Rodríguez, D., Park, B., Cuenca-Estrella, M., Planes, A., Almela, M., Mensa, J., Sánchez, F., Ayats, J., Giménez, M., Saballs, P., Fridkin, S., Morgan, J., Rodríguez-Tudela, J., Warnock, D. & Pahissa, A. (2005). Epidemiology and predictors of mortality in cases of *Candida* bloodstream infection: results from population-based surveillance, Barcelona, Spain, from 2002 to 2003. *J Clin Microbiol*. Vol. (43): 1829-1835.
- Anaissie, E., Vartivarian, S., Abi-Said, D., Uzun, O., Pinczowski, H. & Kontoyiannis, D. (1992). Fluconazole versus amphotericin B in the treatment of hematogenous candidiasis: a matched cohort study. *Am J Med*. Vol. (101): 170-176.
- Arendrup, M., Horn, T. & Frimodt-Moller, N. (2002). *In vivo* pathogenicity of eight medically relevant *Candida* species in an animal model. *Infection*. Vol. (30): 286-291.
- Bautista, M., Boldo, X., Villa-Tanaca L. & Hernández-Rodríguez, C. (2003). Identification of *Candida* spp. by randomly amplified polymorphic DNA and differentiation between *Candida albicans* and *Candida dubliniensis* by direct PCR methods. *J Clin Microbiol*. Vol. (41): 414-420.
- Biasoli, M., Tosello, M. & Magaro, H. (2002). Adherence of *Candida* strains isolated from the human gastrointestinal tract. *Mycoses*. (45): 465-459.
- Bordallo, J. & P. Suarez-Rendueles. (1993). Control of *Saccharomyces cerevisiae* carboxypeptidase S (CPS1) gene expression under nutrient limitation. *Yeast*. Vol. (9): 339-349.
- Bodey, G., Mardani, M., Hanna, H., Boktour, M., Abbas, J., Girgawy, E., Hachem, R., Kontoyiannis, D. & Raad II. (2002). The epidemiology of *Candida glabrata* and *Candida albicans* fungemia in immunocompromised patients with cancer. *Am J Med*. Vol. (112): 380-5.
- Brockert, P., Lachke, S., Srikantha, T., Pujol, C., Galask, R. & Soll, D. (2003). Phenotypic switching and mating type switching of *Candida glabrata* at sites of colonization. *Infect Immun*. Vol. (71): 7109-7118.
- Butler, G., Rasmussen, M., Lin, M., Santos, M., Sakthikumar, S., Munro, C., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J., Agrafioti, I., Arnaud, M., Bates, S., Brown, A., Brunke, S., Costanzo, M., Fitzpatrick, D., de Groot, P., Harris, D., Hoyer, L., Hube, B., Klis, F., Kodira, C., Lennard, N., Logue, M., Martin, R., Neiman, A., Nikolaou, E., Quail, M., Quinn, J., Santos, M., Schmitzberger, F., Sherlock, G., Shah, P., Silverstein, K., Skrzypek, M., Soll, D., Staggs, R., Stansfield, I., Stumpf, M., Sudbery, P., Srikantha, T., Zeng, Q., Berman, J., Berriman, M., Heitman, J., Gow, N., Lorenz, M., Birren, B., Kellis, M. & C. A. Cuomo. (2009). Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature*. Vol. (459): 657-662.

- Byrne, K. & Wolfe, K. (2005). The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* Vol. (15): 1456-1461.
- Campanella, J., Bitincka, L. & Smalley, J. (2003). MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences. *BMC Bioinformatics.* Vol. (4): 29.
- Caro, L., Tettelin, H., Vossen, J., Ram, A., van den Ende, H. & Klis, F. (1997) In silico identification of glycosylphosphatidylinositol-anchored plasma-membrane and cell wall proteins of *Saccharomyces cerevisiae*. *Yeast.* Vol. (13): 1477-1489.
- Castañó, I., Cormack, B. & De Las Peñas, A. (2006). Virulence of the opportunistic pathogen mushroom *Candida glabrata*. *Rev Latinoam Microbiol.* Vol. (48): 66-69.
- Castañó, I., Pan, S., Zupancic, M., Hennequin, C., Dujon, B., & Cormack, B. P. (2005). Telomere length control and transcriptional regulation of subtelomeric adhesins in *Candida glabrata*. *Mol Microbiol.* Vol. (55): 1246-1258.
- Cawley, N., Chino M., Maldonado A., Rodriquez Y., Loh Y. & Ellman, J. (2003). Synthesis and characterization of the first potent inhibitor of yapsin 1. *J Biol Chem.* Vol. (278): 5523-5530.
- Chaffin, W., Lopez-Ribot J., Casanova, M., Gozalbo M. & Martinez J. (1998). Cell wall and secreted proteins of *Candida albicans*: identification, function, and expression. *Microbiol Mol Biol Rev.* Vol. (62): 130-180.
- Chakrabarti, A., Nayak, N. & Talwar, P. (1991). *In vitro* proteinase production by *Candida* species. *Mycopathologia.* Vol. (114): 163-168.
- Copping, V., Barelle, C., Hube, B., Gow, N., Brown, A. & Odds, F. (2005). Exposure of *Candida albicans* to antifungal agents affects expression of *SAP2* and *SAP9* secreted proteinase genes. *J Antimicrob Chemother.* Vol. (55): 645-654.
- Csank, C. & Haynes, K. (2000). *Candida glabrata* displays pseudohyphal growth. *FEMS Microbiol Lett.* Vol. (189): 115-120.
- Dayhoff, M. (1979). *Atlas of protein sequence and structure*, vol. 5, Suppl. 3, National Biomedical Research Foundation, Silver Springs, 978-0912466071, Maryland, p. 353-358.
- Diezmann, S., Cox, C., Schöniar, G., Vilgalys, R. & Mitchell. T. (2004). Phylogeny and evolution of medical species of *Candida* and elated taxa: a multigenic analysis. *J Clin Microbiol.* Vol. (42): 5624-5635.
- Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C., Hofmann, K. & Bairoch, A. (2002). The PROSITE database. *Nucleic Acids Res.* Vol. (30): 235-238.
- Feng, Z., Jiang, B., Chandra, J., Ghannoum, M., Nelson, S. & Weinberg, A. (2005). Human beta-defensins: differential activity against *Candida* species and regulation by *Candida albicans*. *J Dent Res.* Vol. (84): 445-450.
- Fidel, P., Vazquez, J. & Sobel, J. (1999). *Candida glabrata*: review of epidemiology, pathogenesis, and clinical disease with comparison to *C. albicans*. *Clin Microbiol Rev.* Vol (12): 80-96.
- Galtier, N., Gouy, M. & Gautier, C. (1996). SeaView and Phylowin, two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci.* Vol. (12): 543-548.

- Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Botstein, D. & Brown, P. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*. Vol. (11): 4241-4257.
- Ghannoum, M. (2000). Potential role of phospholipases in virulence and fungal pathogenesis. *Clin Microbiol Rev*. Vol. (13): 122-143.
- Gilfillan, G., Sullivan, D., Haynes, K., Parkinson, T., Coleman, D. & Grow, N. (1998). *Candida dubliniensis*: phylogeny and putative virulence factors. *Microbiology*. Vol. (144): 829-838.
- Goodman, J., Winston, D., Greenfield, R., Chandrasekar, P., Fox, B. & Kaizer, H. (1992). A controlled trial of fluconazole to prevent fungal infections in patients undergoing bone marrow transplantation. *N Engl J Med*. Vol. (326): 845-851.
- Helmerhorst, E., Venuleo, C., Beri, A. & Oppenheim, F. (2005). *Candida glabrata* is unusual with respect to its resistance to cationic antifungal proteins. *Yeast*. Vol. (22): 705-714.
- Hicke, L. & Riezman, H. (1996). Ubiquitination of a yeast plasma membrane receptor signals its ligandstimulated endocytosis. *Cell*. Vol. (84): 277-87.
- Hoyer, L., Fundyga, R., Hecht, J., Kapteyn, J., Klis, F. & Arnold, J. (2001). Characterization of Agglutinin-like Sequence Genes From Non-*albicans* *Candida* and Phylogenetic Analysis of the ALS Family. *Genetics*. Vol. (157): 1555-1567.
- Hube, B. (1998). Possible role of secreted proteinases in *Candida albicans* infections. *Rev Iberoam Micol*. Vol. (15): 65-68.
- Hube, B. & J. Naglik. (2001). *Candida albicans* proteinases: resolving the mystery of a gene family. *Microbiology*. Vol. (147): 1997-2005.
- Hube, B., Sanglard, D., Odds, F., Hess, D., Monod, M., Schäfer, W., Brown, A. & Gow, N. (1997). Disruption of each of the aspartyl proteinase genes *SAP1*, *SAP2*, and *SAP3* of *Candida albicans* attenuates virulence. *Infect Immun*. Vol. (65): 3529-3538.
- Hull, C., Raisner, R. & Johnson, A. (2000). Evidence for mating of the "asexual" yeast *Candida albicans* in a mammalian host. *Science*. Vol. (289): 307-310.
- Kalkanci, A., Bozdayi, G., Biri, A., & Kustimur, S. (2005). Distribution of secreted aspartyl proteinase using a polymerase chain reaction assay with *SAP* specific primers in *Candida albicans* isolates. *Folia Microbiol*. Vol. (50): 409-413.
- Joly, S., Maze, C., McCray, P. & Guthmiller, J. (2004). Human betadefensins 2 and 3 demonstrate strain-selective activity against oral microorganisms. *J Clin Microbiol*. Vol. (42): 1024-1029.
- Jones, E. (1991). Three proteolytic system in the yeast *Saccharomyces cerevisiae*. *J Biol Chem*. Vol. (266): 7963-7966.
- Kaur, R., B. & Cormack, B. (2007). A family of glycosylphosphatidylinositollinked aspartyl proteases is required for virulence of *Candida glabrata*. *Proc Natl Acad Sci USA*. Vol. (104): 7628-7633.
- Klionsky, D. & Emr, S. (2000). Autophagy as a regulated pathway of cellular Degradation. *Science*. Vol. (290): 1717-1721
- Krcmery, K. (1999). *Torulopsis glabrata* – an emerging yeast pathogen in cancer patients. *Int J Antimicrob Agents*. Vol. (11): 1-6.
- Krcmery, V., Oravcova, E., Spanik, S., Mrazova-Studena, M., Trupl, J. & Kunova, A. (1998). Nosocomial breakthrough fungaemia during antifungal prophylaxis or empirical antifungal therapy in 41 cancer patients receiving antineoplastic chemotherapy:

- analysis of etiology risk factors and outcome. *J Antimicrob Chemother.* Vol. (41): 373-380.
- Krysan, D., Ting, E., Abeijon, C., Kroos, L. & Fuller, R. (2005). Yapsins are a family of Aspartyl Protease required for cell wall integrity in *Saccharomyces cerevisiae*. *Eukaryot Cell.* Vol. (4): 1364-1374.
- Lachke, S. Joly, S., Daniels, K. & Soll, D. (2002). Phenotypic switching and filamentation in *Candida glabrata*. *Microbiology.* Vol. (148): 2661-2674.
- László Patthy (1999). *Protein Evolution*. Blackwell Science Ltd, ISBN 0-632-04774-7, London.
- Lermann, U. & Morschhauser, J. (2008). Secreted aspartic proteases are not required for invasion of reconstituted human epithelia by *Candida albicans*. *Microbiology.* Vol. (154): 3281-3295.
- Lesage, G., Sdicu, A., Menard, P., Shapiro, J., Hussein, S. & Bussey, H. (2004). Analysis of  $\beta$ -1,3-glucan assembly in *Saccharomyces cerevisiae* using a synthetic interaction network and altered sensitivity to caspofungin. *Genetics.* Vol. (167): 35-49.
- Li, L., Kashleva, H. & Dongari-Bagtzoglou, A. (2007a). Cytotoxic and cytokine inducing properties of *Candida glabrata* in single and mixed oral infection models. *Microb Pathog.* Vol. (42): 138-147.
- Li, L., Redding, S. & Dongari-Bagtzoglou, A. (2007b). *Candida glabrata*, an emerging oral opportunistic pathogen. *J Dent Res.* 86: 204-215.
- Lipke, P. & Ovalle, R. (1998) Cell wall architecture in yeast: new structure and new challenges. *J Bacteriol.* Vol. (180): 3735-3740.
- Loaiza-Loeza, S., Parra-Ortega, B., Bautista-Muñoz, C., Casiano-Rosas, C., Hernández-Rodríguez, C. H. & Villa-Tanaca, L. (2007). The proteolytic system of *Candida dubliniensis*. *Am J Infect Dis.* Vol. (3): 76-83.
- Loaiza-Loeza, S., Parra-Ortega, B., Cancino-Díaz, J., Illades-Aguir, B., Hernández-Rodríguez, C. & Villa-Tanaca, L. (2009). Differential expression of *Candida dubliniensis* secreted aspartyl proteinase genes (*CdSAP1-4*) under different physiological conditions and during the infection of a keratinocytes culture. *FEMS Immunol Med Microbiol.* Vol. (56): 212-22.
- Luo, G. & Samaranayake, L. (2002). *Candida glabrata*, an emerging fungal pathogen, exhibits superior relative cell surface hydrophobicity and adhesion to denture acrylic surfaces compared with *Candida albicans*. *APMIS.* Vol. (110): 601-610.
- Manzano-Gayosso, P., Hernandez-Hernandez, F., Bazan-Mora, E., Mendez-Tovar, L., Gonzalez-Monroy, J. & López-Martínez, R. (2000). Identification and typing of yeast isolates from hospital patients in Mexico City. *Rev Argent Microbiol.* Vol. (32): 1-6.
- MacKay, V., Welch, S., Insley, M., Manney, T., Holly, J., Saari, G. & Parker, M. (1988). The *Saccharomyces cerevisiae* *BAR1* gene encodes an exported protein with homology to pepsin. *Proc Natl Acad Sci USA.* Vol. (85): 55-59.
- Magee, B. & Magee, P. (2000). Induction of mating in *Candida albicans* by construction of MTL $\alpha$  and MTL $\alpha$  strains. *Science.* Vol. (289): 310-313.
- Mayor, S. & Riezman, H. (2004). Sorting GPI-anchored proteins. *Nat Rev Mol Cell Biol.* Vol. (5): 110-120.
- Méan, M., Marchetti, O. & Calandra, T. (2008). Bench-to-bedside review: *Candida* infections in the intensive care unit. *Crit Care.* Vol. (12): 1-9.

- Monod, M., Hube, B., Hess, D. & Sanglard, D. (1998). Differential regulation of *SAP8* and *SAP9* which encode two new members of the secreted aspartyl protease family in *Candida albicans*. *Microbiology*. Vol. (244): 2731-2737.
- Morrow, B., Srikantha, T. & Soll, D. (1992). Transcription of the gene for a pepsinogen, *PEP1*, is regulated by white-opaque switching in *Candida albicans*. *Mol Cell Biol*. Vol. (12): 2997-3005.
- Naglik, J., Albrecht, A., Bader, O. & Hube, B. (2004). *Candida albicans* proteinases and host/pathogen interactions. *Cell Microbiol*. Vol. (6): 915-926.
- Naglik, J., Challacombe, S. & Hube, B. (2003). *Candida albicans* secreted aspartyl proteinases in virulence and pathogenesis. *Microbiol Mol Biol Rev*. Vol. (67): 400-428.
- Naglik, J., Moyes, D., Makwana, J., Kanzaria, P., Tsihlaki, E., Weindl, G., Tappuni, A., Rodgers, C., Woodman, A., Challacombe, S., Schaller, M. & Hube, B. (2008). Quantitative expression of the *Candida albicans* secreted aspartyl proteinase gene family in human oral and vaginal candidiasis. *Microbiology*. Vol. (154): 3266-3280.
- Nikawa, H., Nishimura, H., Yamamoto, T. & Samaranayake, L. (1995). A novel method to study the hyphal phase of *Candida albicans* and to evaluate its hydrophobicity. *Oral Microbiol Immunol*. Vol. (10): 110-114.
- Palmer, G. (2007). Autophagy in the Invading Pathogen. *Autophagy*. Vol. (3): 251-253. Addendum to: Palmer, G., Michelle, N. & Sturtevant, J. (2007). Autophagy in the Pathogen *Candida albicans*. *Microbiology*. Vol. (153): 51-58.
- Palmer, G., Kelly, M. & Sturtevant J. (2005). The *Candida albicans* vacuole is required for differentiation and efficient macrophage killing. *Eukaryot. Cell*. Vol. (4): 1677-1686.
- Parr, C., Keates, R., Bryksa, B., Ogawa, M. & Yada, R. (2007). The structure and function of *Saccharomyces cerevisiae* proteinase A. *Yeast*. Vol. (24): 467-80.
- Parra-Ortega, B., Cruz-Torres, H., Villa-Tanaca, L., Hernández-Rodríguez, C. (2009). Phylogeny and evolution of the aspartyl protease family from clinically relevant *Candida* species. *Mem Inst Oswaldo Cruz, Rio de Janeiro*. Vol. (104): 505-512.
- Pfaller, M. & Diekema, D. (2007). Epidemiology of Invasive Candidiasis: a Persistent Public Health Problem. *Clin Microbiol Rev*. Vol. (20): 133-163.
- Robinson, J., Klionsky, D., Banta, L. & cEmr, S. (1988). Protein sorting in *Saccharomyces cerevisiae*: isolation of mutants defective in the delivery and processing of multiple vacuolar hydrolases. *Mol Cell Biol*. Vol. (8): 4936-4948.
- Roetzer, A., Gregori, C., Jennings, A., Quintin, J., Ferrandon, D., Butler, G. Kuchler, K., Ammerer, G. & Schüller, C. (2008). *Candida glabrata* environmental stress response involves *Saccharomyces cerevisiae* Msn2/4 orthologous transcription factors. *Mol Microbiol*. Vol. (69): 603-620.
- Safdar, A., Van Rhee, F., Henslee-Downey, J., Singhal, S., Mehta, J. & Bone, M. (2001). *Candida glabrata* and *Candida krusei* fungemia after high-risk allogeneic marrow transplantation: no adverse effect of low-dose fluconazole prophylaxis on incidence and outcome. *Bone marrow Transplant*. Vol. (28): 873-8.
- Samaranayake, Y., MacFarlane, T., Samaranayake, L. & Aitchison, T. (1994). The *in vitro* proteolytic and saccharolytic activity of *Candida* species cultured in human saliva. *Oral Microbiol Immunol*. Vol. (9): 229-235.



- Sanglard, D., Hube, B., Monod, M., Odds, F. & Gow, N. (1997). A triple deletion of the aspartyl proteinase genes *SAP4*, *SAP5*, and *SAP6* of *Candida albicans* causes attenuated virulence. *Infect Immun.* Vol. (65): 3539-3546.
- Sanglard, D., Ischer, F., Calabrese, D., Majcherczyk, P. & Bille, J. (1999). The ATP binding cassette transporter gene *CgCDR1* from *Candida glabrata* is involved in the resistance of clinical isolates to azole antifungal agents. *Antimicrob Agents Chemother.* Vol. (43): 2753-2765.
- Scannell, D., Butler, G. & Wolfe, K. (2007). Yeast genome evolution the origin of the species. *Yeast.* Vol. (24): 929-942.
- Schaefer, D., Cote, P., Whiteway, M., & Bennett, R. (2007). Barrier activity in *Candida albicans* mediates pheromone degradation and promotes mating. *Eukaryot cell.* Vol. (6): 907-918.
- Schaller, M., Bein, M., Korting, H., Baur, S., Hamm, G., Monod, M., Beinhauer, S. & Hube, B. (2003). The secreted aspartyl proteinases Sap1 and Sap2 cause tissue damage in an in vitro model of vaginal candidiasis based on reconstituted human vaginal epithelium. *Infect Immun.* Vol. (71): 3227-3234.
- Schaller, M., Mailhammer, R., Grassl, G., Sander, C., Hube, B. & Korting, H. (2002). Infection of human oral epithelia with *Candida* species induces cytokine expression correlated to the degree of virulence. *J Invest Dermatol.* Vol. (118): 652-657.
- Schaller, M., Schäfer, W., Korting, H. & Hube, B. (1998). Differential expression of secreted aspartyl proteinases in a model of human oral candidosis and in patient samples from the oral cavity. *Mol Microbiol.* Vol. (29): 605-615.
- Sobel, J. (2006). The emergence of non-*albicans* *Candida* species as causes of invasive candidiasis and candidemia. *Curr Infect Dis Rep.* Vol. (8): 427-433.
- Srikantha, T., Lachke, S. & Soll, D. (2003) Three mating type-like loci in *Candida glabrata*. *Eukaryot Cell.* Vol. (2): 328-340.
- Szklarczyk, R., & Heringa, J. (2004). Tracking repeats using significance and transitivity. *Bioinformatics.* Vol. (20): 311-317.
- Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol.* Vol. (24): 1596-1599.
- Taylor, B., Staib, P., Binder, A., Biesemeier, A., Sehnal, M., Rollinghoff, M., Morschhauser, J. & Schroppel, K. (2005). Profile of *Candida albicans*-secreted aspartic proteinase elicited during vaginal infection. *Infect Immun.* Vol. (73): 1828-1835.
- Teichert, U., Mechler, B., Muller, H. & Wolf D. (1989). Lysosomal (vacuolar) proteinases of yeast are essential catalysts for protein degradation, differentiation, and cell survival. *Rev. Microbiol.* Vol. (6): 2500-2510.
- Trick, W., Fridkin, S., Edwards, J., Hajjeh, R. & Gaynes, R. (2002). Secular trend of hospital-acquired candidemia among intensive care unit patients in the United States during 1989-1999. *Clin Infect Dis.* Vol. (35): 627-630.
- Wei, X., Rogers, H., Lewis, M., & Williams, D. (2011). The role of the IL-12 cytokine family in directing T-cell responses in oral candidosis. *Clin Dev Immunol.* Vol. (2011): 1-10.
- White, T. & Agabian, N. (1995). *Candida albicans* secreted aspartil proteinases: isoenzyme pattern is determined by cell type, and levels are determined by enviromental factors. *J Bacteriol.* Vol. (177): 5215-5221.

- White, T., Miyasaki S. & Agabian, N. (1993). Three distinct secreted aspartyl proteinases in *Candida albicans*. *J Bacteriol.* Vol. (175): 6126-6133.
- Wolfe, K. (2006). Comparative genomics and genome evolution in yeasts. *Philos Trans R Soc Lond B Biol Sci.* Vol. (361): 403-412.
- Wünschmann, J., Beck, A., Meyer, L., Letzel, T., Grill, E. & Lenzian, K. (2007) Phytochelatins are synthesized by two vacuolar serine carboxypeptidases in *Saccharomyces cerevisiae*. *FEBS Lett.* Vol. 17:1681-7.

## Clues to Evolution of the SERA Multigene Family in the Genus *Plasmodium*

Nobuko Arisue, Nirianne M. Q. Palacpac,  
Kazuyuki Tanabe and Toshihiro Horii  
*Research Institute for Microbial Diseases, Osaka University  
Japan*

### 1. Introduction

Malaria, one of the most serious infectious diseases prevalent in the tropics, is caused by the genus *Plasmodium*. Despite considerable global efforts to control this parasitic disease at least 90% of deaths still occur in sub-Saharan Africa (WHO, 2010). The rising threat of drug-resistant parasites, together with key interventions dependent on the use of a limited class of insecticides, underscore the fragility of malaria control. A better understanding of the parasite biology is required to gain insights for cost-effective tools or strategies including malaria vaccines and new antimalarial drugs that can be instrumental for sustained control, if not elimination, of malaria.

Malaria parasites comprise a diverse group of over 200 *Plasmodium* species that infect mammals, birds and reptiles (Levine, 1988). Each *Plasmodium* species exhibits a restricted host range, such that primate parasite cannot infect rodent, bird or reptile hosts. To find genomic factors that determine host range transcends the interest of malaria researchers. It is essential for control and conservation of wildlife. Recently, genome projects on several *Plasmodium* species from different hosts have been completed (Gardner et al., 2002; Carlton et al., 2002, 2008; Pain et al., 2008), with gene information available in the public database (<http://plasmodb.org/plasmo/>). By comparing the genomes from different species we obtain basic information at the molecular level on how *Plasmodium* has evolved and allows us to infer the function of genes and noncoding regions in the genome. One of the prominent features of *Plasmodium* genomes is the presence of various unique multigene families. Multigene families are a group of related genes that are presumed to share a common ancestor and are derived from each other by duplication and subsequent divergence. One such example, the largest family identified so far in human, primate and rodent malaria, is the *Plasmodium* interspersed repeat, *pir* (Janssen et al., 2004). The *pir* gene family members are highly species-specific, suggesting evolution of lineage-specific immune evasion mechanisms. *P. falciparum* var gene family is by far the best documented multigene family of the most virulent human malaria parasite. Products of var genes appear on the surface of infected erythrocytes and are involved in antigenic variation to evade host immunity. Other species-specific gene families encode proteins involved in host cell invasion, e.g. rhoptry proteins and parasite surface antigens, merozoite surface protein-3 and -7. There are also examples of families with few gene members. In sharp contrast to several hundreds of tandem arrayed rRNA gene family members in other eukaryotes,

*Plasmodium* has 4-7 gene units physically separated in the genome (Nei & Rooney, 2005; Carlton et al., 2008). Thus, *Plasmodium* possesses unique multigene families with distinctive evolutionary conundrums.

For more than 10 years after the first description of a gene family member, the existence of the *Plasmodium* serine repeat antigen (SERA) multigene family has been overlooked. Serine repeat antigen family proteins share homology with the papain family of cysteine proteases (Kiefer et al., 1996; Gor et al., 1998; Bourgon et al., 2004; Arisue et al., 2007, 2011). Almost all SERA genes are clustered in a head-to-tail manner and the number of SERA genes in the clustered region varies among parasite species (Bourgon et al., 2004; McCoubrie et al., 2007; Arisue et al., 2007, 2011). This leads us to infer that gene duplication occurred repeatedly during evolution. Some SERA genes were confirmed to play essential role(s) in the parasite life cycle (Miller et al., 2002; Aly & Matuschewski, 2005; McCoubrie et al., 2007). In addition, a gene family member in *P. falciparum*, SERA5, is a vaccine candidate now on phase Ib clinical trial (Horii et al., 2010). Two observations promise SERA5 as a vaccine candidate: (1) SERA genes are not differently expressed like other antigen encoding gene families such as var and rifin that show antigenic variation to evade host immune response (Aoki et al., 2002; Miller et al., 2002; Palacpac et al., 2006; Schmidt- Christensen et al., 2008; Putrianti et al., 2010; Arisue et al., 2011); and (2) *P. falciparum* SERA5 is less polymorphic (Fox et al., 1997; Morimatsu et al., 1997; Liu et al., 2000) than other vaccine candidate genes such as merozoite surface protein 1 (McBride et al., 1985) and apical membrane protein 1 (Polley et al., 2003; Cortés et al., 2003). These characteristics are indeed appealing and show the unique biological features of *Plasmodium* SERA. Here, we summarize current reports and our recent findings to understand the evolution of the SERA gene family.

## 2. SERA gene repertoires in *Plasmodium* species

We refer briefly to the research history of the SERA multigene family: (i) the identification of SERA5 and the proteolytic processing of the protein; (ii) the discovery of the gene family following chromosome 2 sequencing of the *P. falciparum* genome; (iii) currently known SERA genes from different species; and (iv) the resulting analyses of the multigene family in various malaria parasites.

### 2.1 Research history of the SERA multigene family

SERA was first found in *P. falciparum* as an abundant, exported, soluble late-trophozoite to schizont stage protein (Perrin et al., 1984). The protein, independently isolated by different groups, was described under various names as Pf140 (Perrin et al., 1984), p113 (Chulay et al., 1987), p126 (Deplace et al., 1987) or SERP (Knapp et al., 1989). All identified a gene with a long stretch of repeated serine residues in the N-terminal domain to which the family owes its name (Bzik et al., 1988). At the central domain, SERA possess a motif which align with two active site-determining regions of cysteine proteinases. The secreted protein was described to accumulate in the parasitophorous vacuole, and released into the culture medium at the time of schizont rupture. Notably, before the sequence of *P. falciparum* Chromosome 2 was opened, Knapp et al. (1991) discovered a SERA homolog (*serp H*) lacking the characteristic serine homopolymer, and subsequently Fox and Bzik (1994) reported SERA as one of three consecutive series of homologous genes. The complete genome sequence of *P. falciparum* revealed that SERA belongs to a multigene family (Gardner et al., 1998). The originally described SERA gene was renamed SERA5 according to

the gene arrangement order in *P. falciparum* (Aoki et al., 2002; Miller et al., 2002). It is interesting to note that SERA5 is the only member with repeated serine residues among nine gene members. The characteristic of the family is not the richness in serine residues but motifs that generate the framework of a cysteine protease. SERA homologs were identified in other *Plasmodium* species. Kiefer et al. (1996) found five SERA genes from another human parasite, *P. vivax* and Gor et al. (1998) identified three SERA genes from the rodent parasite of *P. vinckei*. Completed or ongoing genome projects of eight *Plasmodium* species: two human malaria parasites, *P. falciparum* and *P. vivax*; chimpanzee parasite *P. reichenowi*; macaque parasite *P. knowlesi*; three rodent parasites *P. berghei*, *P. yoelii* and *P. chabaudi*; and avian parasite *P. gallinaceum*, confirmed the gene organization and allowed phylogenetic analysis of the SERA gene family (Burgon et al., 2005; Arisue et al., 2007). In addition, Arisue et al. (2011) newly identified SERA genes in 11 *Plasmodium* species that further elaborate the genome organization of the gene family.

## 2.2 Processing of *P. falciparum* SERA5

The *in vitro* observation that *P. falciparum* SERA5 was released into the culture supernatant at the time of schizont rupture/merozoite release corresponds to its specific processing into several polypeptides. The full-length 120 kDa precursor accumulates in the parasitophorous vacuole during late trophozoite and schizont stages. As shown in Fig. 1., during the course of schizont rupture/merozoite release, SERA is proteolytically processed into a 47 kDa N-terminal (P47), a 50 kDa central (P50), an 18 kDa C-terminal (P18) and a 6 kDa domain (Delplace et al., 1987, 1988; Debrabant et al., 1992; Li et al., 2002a). The N-terminal P47 fragment is further processed into two 25 kDa fragments (P25n and P25c) in some allelic types (Li et al., 2002a). P47 is linked with the C-terminal P18 via disulfide bond that is localized at the merozoite surface that is localized at the merozoite surface (Delplace et al., 1987; Li et al., 2002a; Okitsu et al., 2007). The proteolytic processing is mediated by subtilisin-like serine protease subtilase 1 or SUB1 (Yeoh et al., 2007). Inhibition of SERA maturation blocks parasite egress from the host erythrocyte (Li et al., 2002b; Yeoh et al., 2007).

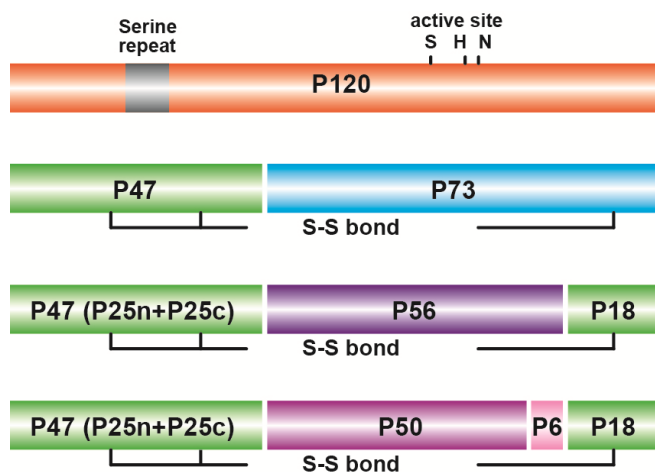


Fig. 1. Processing of *P. falciparum* SERA5 during parasite egress from host erythrocyte.

### 2.3 SERA genes in the database

The discovery of the SERA gene family in *P. falciparum* sparked the subsequent identification of a number of SERA genes in different *Plasmodium* species. Currently known SERA genes in the PlasmoDB database (<http://plasmodb.org/plasmo/>) are summarized in Table 1 and accession numbers of SERA genes found in the public database (NCBI, <http://www.ncbi.nlm.nih.gov/>) are summarized in Table 2.

We opted not to list SERA genes of *P. reichenowi* and *P. gallinaceum* in either Table, but for our analysis, their SERA gene sequences were assembled from various reads in the partial genome shotgun database of *Plasmodium* at The Sanger Institute. Blast programs in the following web sites were used to search SERA coding reads for: *P. reichenowi*: [http://www.sanger.ac.uk/cgi-bin/blast/submitblast/p\\_reichenowi](http://www.sanger.ac.uk/cgi-bin/blast/submitblast/p_reichenowi); and *P. gallinaceum*: [http://www.sanger.ac.uk/cgi-bin/blast/submitblast/p\\_gallinaceum](http://www.sanger.ac.uk/cgi-bin/blast/submitblast/p_gallinaceum)). Identified SERA gene sequences could be referred to in Arisue et al. (2007).

### 2.4 Characteristic features of the SERA multigene family

Almost all SERA genes were aligned in a tandem cluster between the conserved hypothetical protein gene and the iron-sulfur assembly protein gene. According to genetic background, SERA genes can be categorized into Groups I to IV (Arisue et al., 2007, 2011). Characteristic features of each Group are summarized in Fig. 2. Group I to Group III SERA genes possess the protease motif that includes an active site cysteine residue, in contrast to Group IV SERA genes where the cysteine residue is replaced by a serine (Bourgon et al., 2005; Arisue et al., 2007, 2011). The mRNA transcription and/or protein expression of Group I SERA genes were observed in the mosquito vector, while those of Group II to Group IV SERA genes were observed in the vertebrate host (Ali & Matuschewski, 2005; Arisue et al., 2007, 2011; Putrianti et al., 2010). The difference in the gene repertoire among species is due to the number of Group IV SERA genes. SERA gene repertoires are summarized in Table 3.

Species (strain)	<i>P. vivax</i> (Sall)	<i>P. falciparum</i> (3D7)	<i>P. knowlesi</i> (H)	<i>P. berghei</i> (ANKA)	<i>P. yoelii</i> (17XNL)	<i>P. chabaudi</i> (AS)
SERA1	PVX_003850	PFB0360C	PKH_041200	PB000108.03.0	PY00291	PCAS_030730
SERA2	PVX_003845	PFB0355C	PKH_041210	PB107093.00.0	PY00292	PCAS_030720
SERA3	PVX_003840	PFB0350C	PKH_041230	PB000107.03.0	PY00292	PCAS_030710
SERA4	PVX_003835	PFB0345C	PKH_041250	PB000352.01.0	PY02062 PY00294	PCAS_030700
SERA5	PVX_003830	PFB0340C	PKH_041260	PB000649.01.0	PY02063	PCAS_030690
SERA6	PVX_003825	PFB0335C	PKH_041270			
SERA7	PVX_003820	PFB0330C				
SERA8	PVX_003810	PFB0325C				
SERA9	PVX_003805	PFI0135C				
SERA10	PVX_003800					
SERA11	PVX_003795					
SERA12	PVX_003790					

\*The gene region of PY02062 and PY00294 was re-annotated and used as *P. yoelii* SERA4.

Table 1. GeneID of SERA genes in the PlasmoDB (<http://plasmodb.org/plasmo/>).



Species	Natural host	Number of SERA gene in each group				
		I	II	III	IV	Degenerate*
<i>P. falciparum</i>	human	1	1	1	6	0
<i>P. vivax</i>	human	1	1	1	9	2
<i>P. malariae</i>	human	1	1	1	7	3
<i>P. ovale</i>	human	1	1	1	4	1
<i>P. knowltoni</i>	human/ macaque	1	1	1	3	2
<i>P. cynomolgi</i>	macaque	1	1	1	8	3
<i>P. coatneyi</i>	macaque	1	1	1	4	3
<i>P. fragile</i>	macaque	1	1	1	2	3
<i>P. simiovale</i>	macaque	1	1	1	6	3
<i>P. fieldi</i>	macaque	1	1	1	6	3
<i>P. inui</i>	macaque	1	1	1	4	5
<i>P. hylobati</i>	gibbon	1	1	1	4	1
<i>P. berghei</i>	rat	1	1	1	2	0
<i>P. yoelii</i>	rat	1	1	1	2	0
<i>P. chabaudi</i>	rat	1	1	1	2	0
<i>P. gallinaceum</i>	bird	2			1	0
<i>P. gonderi</i>	mangabey, guenon	1	1	1	6	?
<i>P. reichenowi</i>	chimpanzee	1	1	1	5?	?
<i>P. vinckei</i>	rat	?	1	1	1?	?

Table 3. The number of SERA genes that belong to each group from several *Plasmodium* species. 'Degenerate'(\*) denotes defective gene copies, i.e., the total number of pseudogenes, truncated gene and gene fragments found.

## 2.5 Primary structure of SERA molecules and genes

Schematic representation of SERA gene structure is shown in Fig. 3A. Group I SERA genes code for around 700 amino acids. They share a similar six exon and five intron structure, except for *P. falciparum* SERA8 and *P. vivax* SERA12 that both lack one intron. Group II to Group IV SERA genes code for about 1000 amino acid residues, and similar to Group I, share a common four exon/three intron structure with few exceptions. All SERA genes have the structural context of cysteine proteinases, however, it is important to note that the canonical Cys His Asn triad of the active proteinase is not present in all. The relatively small number of amino acid residues in Group I SERA resulted to a shorter N-terminal region when compared to Group II to IV SERA. Multiple amino acid sequence alignments revealed the consensus primary structure of SERA, which consists of six putative domains shown in Fig. 3B.

Amino acid sequences of the putative pro-enzyme and enzyme domains are remarkably conserved, but extensive sequence variations are found in variable domains 1 and 2. In the C-terminal cysteine rich conserved domain, seven cysteine residues are perfectly conserved in all SERA genes.

The pro-enzyme and enzyme domains of *P. falciparum* SERA5 was identified by functional genetic and structural analyses (Hodder et al., 2003, 2009). These domains, corresponding to



P50 in Fig. 1, are flanked by the reported SUB1 cleavage sites (Yeoh et al., 2007). The consensus sequence of the cleavage site is (Val/Leu/Ile)-Xaa-(Gly/Ala)-Paa, in which Xaa is any amino acid residue and Paa is any non-polar residue except for Leu (Yeoh et al., 2007). This consensus sequence is well conserved with slight modifications in all Group II to IV *Plasmodium* SERA genes which we have analyzed. *In vitro*, *P. falciparum* SERA4 (Group IV) and SERA6 (Group II) were cleaved by recombinant PfSUB1 (Yeoh et al., 2007). Peculiarly, Group I SERA genes lack most of the N-terminal variable domain 1 and SUB1 cleavage sites.

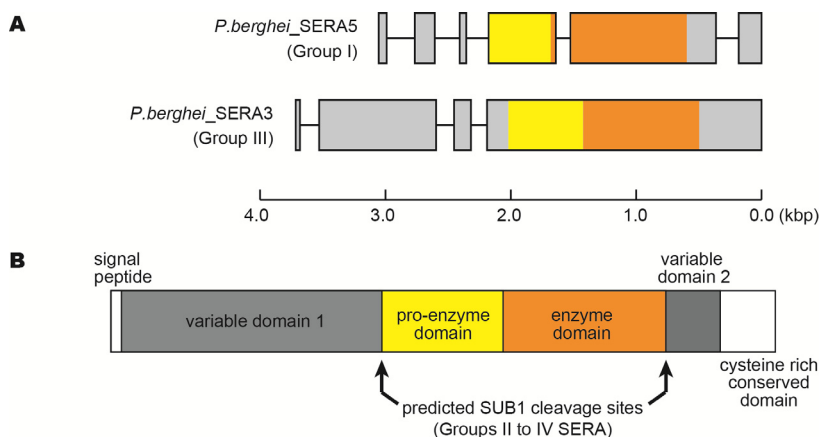


Fig. 3. Schematic representation of the SERA gene structure (A) and their putative domain organization (B).

### 3. Phylogenetic relationships of SERA genes

The categorization of SERA genes into four groups, Group I to IV, was based on phylogenetic analysis (Arisue et al., 2007, 2011). SERA amino acid sequences from 18 *Plasmodium* species were aligned using CLUSTAL W program (<http://clustalw.ddbj.nig.ac.jp/top-j.html>) under default options with manual corrections. Unambiguously aligned amino acid positions corresponding to the putative pro-enzyme domain, enzyme domain and cysteine rich conserved domain were selected and used for the phylogenetic analysis. Maximum likelihood tree was inferred using the PROML program in PHYLIP version 3.69 (Felsenstein, 1996). Except for the number of genes and number of amino acid sequences included in the analysis, the same method was used to infer the phylogenetic tree shown in Fig. 4 and 5.

A simplified maximum likelihood tree inferred from 134 SERA genes with 392 amino acid positions is shown in Fig. 4. Bootstrap proportion values were placed only on the common ancestor branch of each group. The monophyletic grouping of Group I SERA genes was supported with a bootstrap value of 100%. The long internal branch separating Group I from Groups II to IV suggests that the root of the tree is located on the branch leading to the common ancestor of Group I SERA genes. It is thus likely that Group I genes have appeared early in the evolution of the SERA gene family. *P. gallinaceum* SERA1 branches at the common ancestor of Group II to IV, suggesting that gene duplication events which produced Groups II, III and IV likely occurred after the divergence of *P. gallinaceum* from the common ancestral lineage of *Plasmodium*.

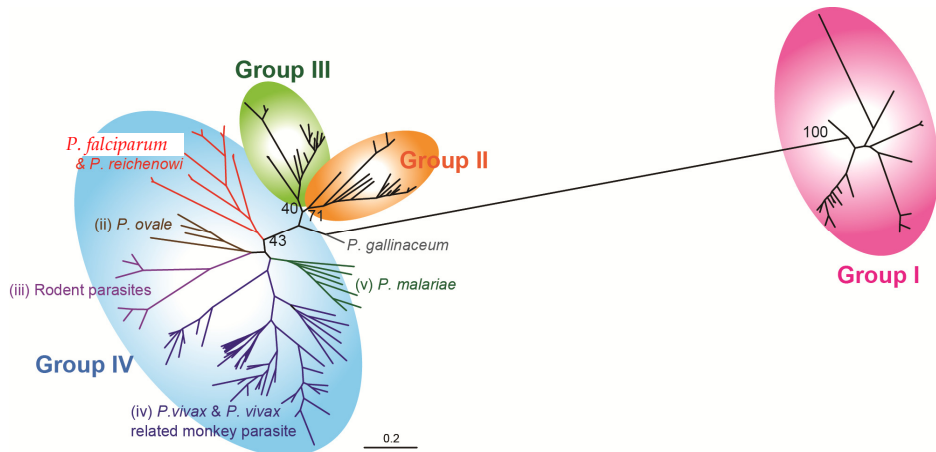


Fig. 4. Phylogenetic tree based on the alignment of 134 SERA genes from 18 *Plasmodium* species.

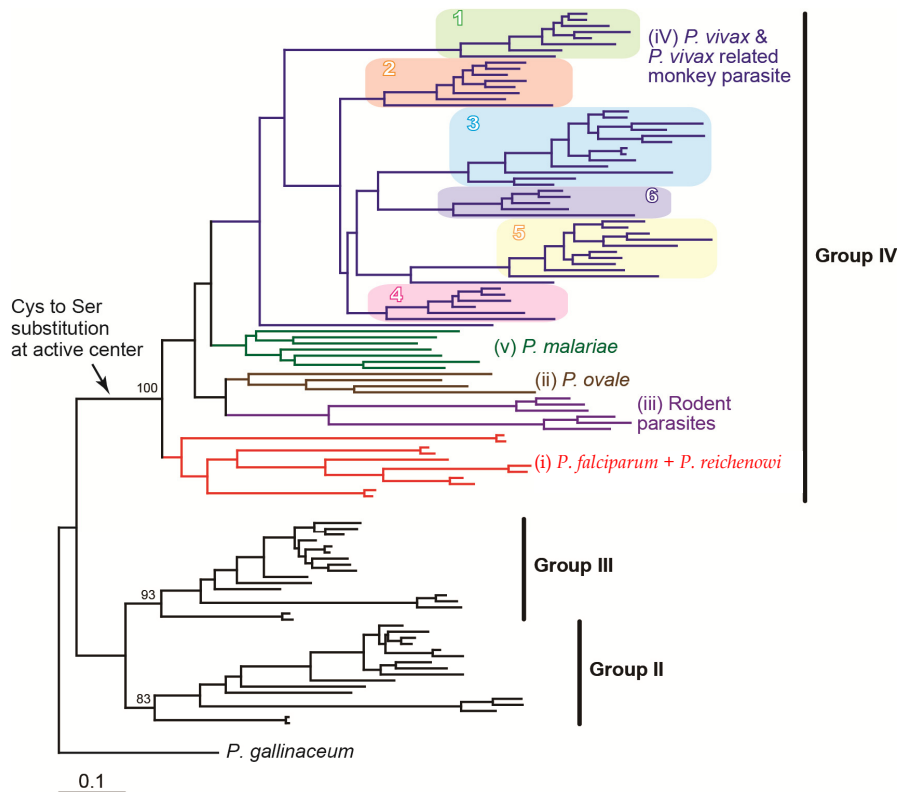


Fig. 5. Phylogenetic tree based on the alignment of 115 SERA genes from 18 *Plasmodium* species.

Group IV SERA genes which diverged after the cysteine to serine substitution in the catalytic site were further divided to five monophyletic sub-Groups: (i) *P. falciparum* and *P. reichenowi*, (ii) *P. ovale*, (iii) rodent *Plasmodium* species, (iv) *P. vivax* and *P. vivax*-related monkey parasite species, and (v) *P. malariae*. These indicate that genes have duplicated independently in each of the sub-group lineages. To increase the resolution of the tree, the long branched Group I genes were excluded from the dataset and the maximum likelihood tree was re-constructed from 115 SERA genes categorized into Group II to Group IV. The resultant tree is shown in Fig. 5.

Interestingly, Group IV SERA genes of *P. vivax* and *P. vivax*-related monkey parasites (10 species) were further categorized into six orthologous gene groups, namely, Clade 1 to Clade 6; and each clade has 5 (Clade 6) to 10 (Clade 5) parasite species. The number of SERA genes analyzed varied from 5 (*P. fragile*) to 12 (*P. vivax*). This does suggest that a common ancestor of *P. vivax* and related monkey malaria parasites had at least 6 SERA genes of Group IV; and that gene duplications and gene deletions occurred in each lineage. Orthologous relationships between SERA gene members and their relative arrangement are shown in Fig. 6.

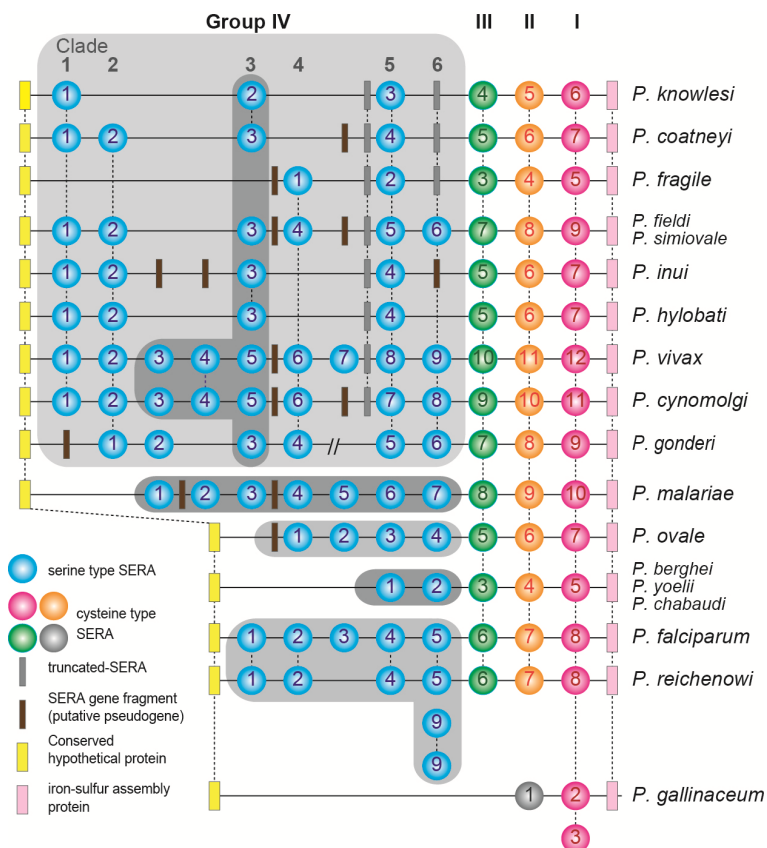


Fig. 6. Organization and phylogenetic relationship of SERA gene in 18 *Plasmodium* species.

In addition to several gene members characterizing Group IV, there are multiple SERA gene fragments and pseudogenes containing multiple stop codons. Taken together, these extensive gene duplications, gene deletions as well as pseudogenization/truncation are evident only in the serine type SERA gene (Group IV) of *P. vivax* and related monkey malaria parasites.

#### 4. Transcription analyses of SERA genes

Transcription analyses revealed, likewise, some discordance among *Plasmodium* species. Transcription profile of the SERA gene family was analyzed first by Aoki et al. (2002) in *P. falciparum*. Genes were most actively transcribed at the late trophozoite to schizont stages of the parasite with SERA5 predominantly transcribed among the family (Fig. 7.).

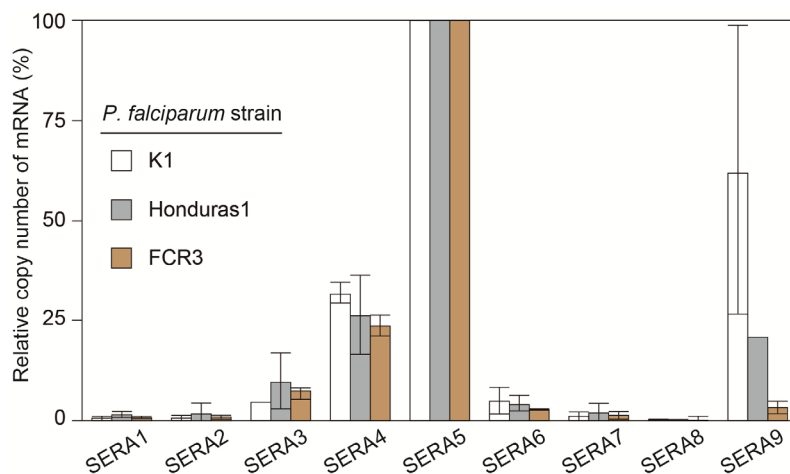


Fig. 7. The relative abundance of mRNA for each *P. falciparum* SERA gene during late trophozoite to schizont stages of the parasite.

Similar to *P. falciparum*, multiple number of SERA genes were transcribed at late trophozoite to schizont stages when transcription analysis was done for the SERA gene family in other *Plasmodium* parasites: human parasite *P. vivax* (Palacpac et al., 2006); rodent parasite *P. berghei* (Arisue et al., 2011); and three monkey parasites *P. knowlesi*, *P. cynomolgi* and *P. coatneyi* (Arisue et al., 2011). A representative summary of SERA gene transcription analysis is shown in Fig. 8. In malaria parasite species infecting humans, one of Group IV SERAs of *P. falciparum* (SERA5) and *P. vivax* (SERA4) showed the highest transcription level among other gene members. In the rodent malaria parasite *P. berghei*, Group III SERA gene (SERA3) was predominantly expressed. In three monkey malaria parasites, the abundantly expressed genes are members of Group IV Clade 3: SERA3 and SERA5, in *P. cynomolgi*; SERA3 in *P. coatneyi*; and SERA2 in *P. knowlesi*. These results show that SERA genes were differently expressed between rodent and primate parasites. Based on the malaria parasite mitochondrial genome, *P. falciparum* belongs to the primate parasite group 1 lineage whereas *P. vivax* and the three macaque parasites belong to primate parasite lineage 2. Phylogenetic analysis showed that these two lineages are not closely related; and the rodent parasite lineage is positioned between them (Hayakawa et al., 2008). Note, however, that

both primate lineages showed similar transcription pattern of SERA gene which might suggest a possible relationship between SERA function and host specificity.

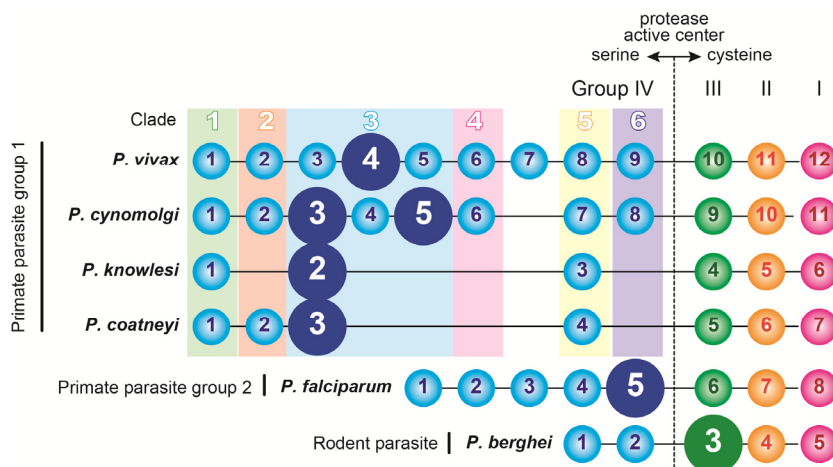


Fig. 8. Representation summary of SERA gene transcription analyses in primate and rodent *Plasmodium* spp. Abundantly transcribed SERA genes are differentiated using enlarged solid circles.

## 5. Duplication in the multigene family

In general, duplicated genes undergo either (i) concerted evolution or (ii) birth-and-death evolution (Nei & Rooney, 2005). In the concerted evolution model, all members of a gene family evolve as a unit rather than independently. When a mutation occurs in a gene, mutation spreads to all other gene members by unequal crossover or gene conversion. As a result, all members of the gene family show identical sequence to each other. The evolution of rRNA multigene families in vertebrates is a classic example of concerted evolution. Analysis of MHC genes in mammals (Hughes & Nei, 1989; Nei et al., 1997; Nei & Hughes, 1992), other immune system related genes (Hughes & Nei, 1990; Ota & Nei, 1994) and disease-related genes (Zhang et al., 2000) show a quite different evolutionary pattern. The birth-and-death evolution model was proposed to explain differential duplication/independent diversification processes that result to subsequent loss or maintenance of genes in a multigene family. Thus, some duplicated genes are maintained in the genome for a long time while others are deleted or became pseudogenes through deleterious mutations. This model applies to rRNAs of *Plasmodium* species in marked contrast to the concerted evolution of rRNAs in most organisms. The model aptly explains the observation that rRNA genes in *Plasmodium* were structurally and functionally distinct (Rooney, 2004; Nishimoto et al., 2008).

The observed gene duplication and gene deletion found in the *Plasmodium* SERA genes are clearly in concordance with the birth-and-death model, although traits of gene conversion are detected in a few of Group IV SERA genes. The birth-and-death model has, likewise, been recently proposed for gene duplication/gene deletion of merozoite surface protein 7, an immune target parasite surface antigen gene (Garzón-Ospina et al., 2010). It, thus, seem

most probable that diversification of *Plasmodium* SERA multigene family was also driven by the birth-and-death evolution. Inferred gene duplication events in the evolution of the *Plasmodium* SERA gene family are shown in Fig. 9.

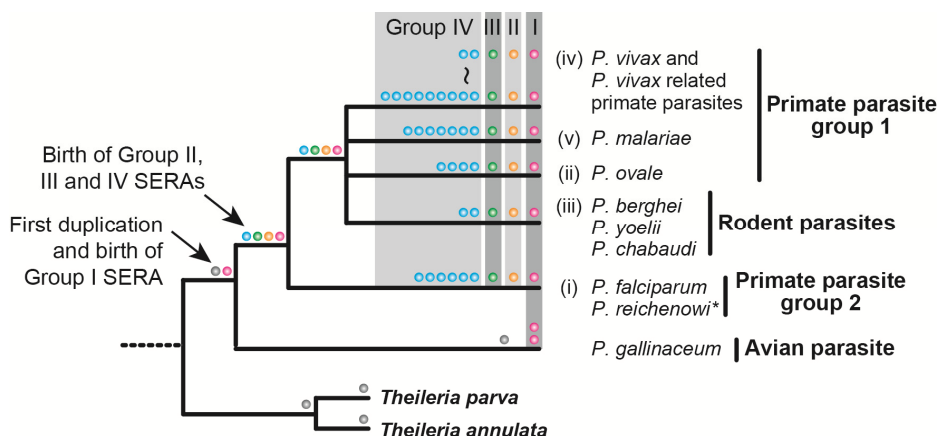


Fig. 9. Inferred gene duplication events in the evolution of the *Plasmodium* SERA gene family. Asterisk in *P. reichenowi* denotes that the SERA gene number in this species is tentative.

*Theileria* is the only other genus to possess a SERA homolog gene. The apicomplexan parasite is closely related to *Plasmodium*. Sequence similarity search against *Theileria* genome database at The Sanger Institute identified a single gene from both *T. parva* and *T. annulata* which has similarity with cysteine-type SERA (Arisue et al., 2007; McCoubrie et al., 2007). Based on Fig. 9, as all *Plasmodium* species have multiple SERA genes, we likely infer that the first duplication occurred at the common ancestor lineage of *Plasmodium*. Because every *Plasmodium* species has Group I SERA gene, the first duplication event is from Group I SERA gene. The duplication events which gave rise to Group II and IV SERA genes occurred after the divergence of *P. gallinaceum* from the branch leading to a common ancestral species of other *Plasmodium* species since *P. gallinaceum* has no Group II to IV SERA gene. The rest of the 17 *Plasmodium* species might have diverged into five lineages of (i) *P. falciparum* and *P. reichenowi*, (ii) *P. ovale*, (iii) rodent *Plasmodium* species, (iv) *P. vivax* and *P. vivax*-related monkey parasite species, and (v) *P. malariae*, and duplications of Group IV SERA genes occurred independently on each lineage. In addition, gene deletions as well as pseudogenization/truncation occurred frequently in *P. vivax* and *P. vivax*-related primate parasite lineage.

## 6. Conclusion and open issues in SERA study

Multigene families are believed to provide an organism with a set of related genes that allow fine tuning of its biological function with possibly different temporal or topologic expression patterns. SERA gene duplications during *Plasmodium* evolution generated four types of SERA genes: Group I to Group IV. The speculated function of SERA during the parasite life cycle is summarized in Fig. 10.

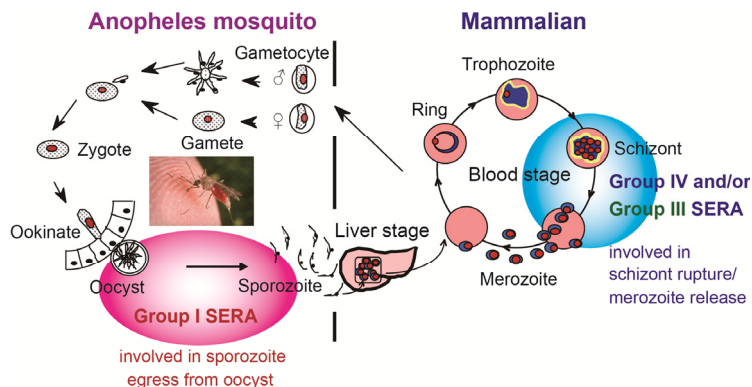


Fig. 10. The life cycle of the malaria parasite and inferred role(s) of SERA.

Group I SERA gene, bearing the canonical cysteine, was shown to be transcribed in the oocyst stage and its gene product was a protease required for sporozoite egress from the oocyst (Aly & Matuschewski, 2005). Group III SERA gene was suggested to play an essential role during schizont rupture/merozoite release in the mammalian host (Yeoh et al., 2007; Arastu-Kapur et al., 2008; Putrianti et al., 2010). Group IV SERA genes bear the characteristic replacement of active-site cysteine to a serine residue. Perhaps importantly, only mammalian parasites have Group IV SERA gene; and Group IV SERA gene of primate parasite was suggested to play an essential role in schizont rupture/merozoite release together with Group III SERA gene (Yeoh et al., 2007; Arisue et al., 2011). The duplication of Group IV SERA gene occurred particularly frequent in two evolutionarily distinct primate lineages and it is intriguing to assume that duplications of SERA genes were associated with host range expansion.

The study of the SERA gene family points to its unique features reinforcing the importance of investigating other uncharacterized gene families of *Plasmodium* to further understand the evolutionary history and biology of this harmful parasite. Many questions still remain in the analysis of SERA. SERA genes are thought to be subject to birth-and-death evolution, and thus, a pattern of interspecific gene clustering is expected to characterize the SERA family whereby functional genes are maintained in the genome for a long time and others are deleted or become non-functional. Group I and Group III SERA genes are highly conserved in *Plasmodium* species. For Group II SERA genes, although maintained among *Plasmodium* species with significant sequence similarity, no function has yet been predicted. Gene disruption studies with Group II SERA gene of *P. berghei* showed no apparent phenotypic change (Arisue et al., unpublished data). Group II is similar to Group I and Group III in being a cysteine-type SERA gene which has been suggested to have proteolytic activity to cleave host membrane structure (Aly & Matuschewski, 2005; Yeoh et al., 2007). The papain-like cysteine protease motif in its amino acid sequence suggests the possibility that Group II SERA act as a protease sometime in the parasite life cycle. Parasite egress from the host cell is an important process that remains poorly understood.

*P. falciparum* SERA5 is a vaccine candidate molecule now on clinical trial in Uganda (Horii et al., 2010). Serum antibodies against the N-terminal domain of *P. falciparum* SERA5 in individuals living in malaria endemic area protect infants from clinical malaria and inhibit *in vitro* parasite growth (Okech et al., 2001, 2006; Aoki et al., 2002; Horii et al., 2010). During



blood stage growth, all SERA gene family member of *P. falciparum* are transcribed most actively at trophozoite and schizont stages. SERA5 is the most abundantly expressed gene family member, with expression levels estimated to be approximately 0.5-1.5% of the whole mRNA at schizont stage (Aoki et al, 2002). However, sero-positivity rate against the N-terminal domain of *P. falciparum* SERA5 was observed to be relatively low (Fig. 11.; Aoki et al., 2002; Horii et al., 2010).

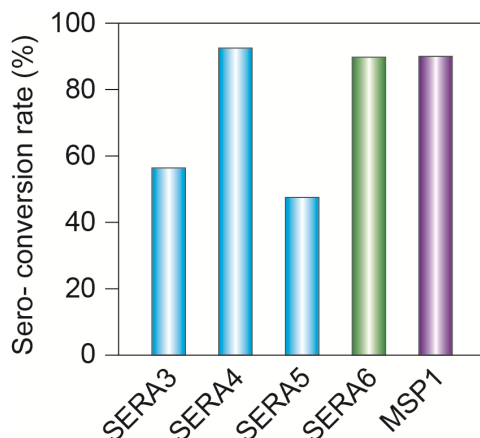


Fig. 11. Relative sero-conversion rates of *P. falciparum* SERA3 to SERA6 and merozoite surface protein 1 (MSP1) in a malaria endemic area of Uganda.

Since the SERA gene family does not show antigenic variation to evade host immune response (Fig. 7.), there may possibly be another mechanism of host parasite evasion/molecular mimicry/interference or competition.

The host range or host specificity of *Plasmodium* is believed to be restricted, although, primate malaria parasites generally infect multiple hosts. For example, it has been reported that, *P. knowlesi* and *P. cynomolgi* have the ability to infect a wide variety of macaques and human (Coatney et al., 1971); additionally, two human parasites *P. malariae* and *P. ovale* have been detected in chimpanzees (Hayakawa et al., 2009; Duval et al., 2009). It may be probable that duplications of Group IV SERA genes that occurred frequently in both primate parasite lineages may be associated with host range expansions. To date, no experimental support lends credence to this speculation.

As described above, the molecular function of SERA genes in each group, the relationship of immune evasion mechanism and the SERA gene family, and the association of host range with Group IV SERA genes remain important issues that needs to be addressed. The importance of SERA genes in parasite egress and their role in host-parasite interactions serve to propel further studies in understanding this multigene family.

## 7. Acknowledgment

This work was supported by **KAKENHI (18073013 and 20390120)** from the Japanese Ministry of Education, Science, Sports, Culture and Technology.



## 8. References

- Aly, ASI. & Matuschewski K. (2005) A malarial cysteine protease is necessary for *Plasmodium* sporozoite egress from oocysts, *The Journal of Experimental Medicine*, Vol. 202, No. 2, pp. 225-230.
- Aoki, S.; Li J, Itagaki S et al. (2002) Serine repeat antigen (SERA5) is predominantly expressed among the SERA multigene family of *Plasmodium falciparum*, and the acquired antibody titers correlate with serum inhibition of the parasite growth, *The Journal of Biological Chemistry*, Vol. 277, No 49, pp. 47533-47540.
- Arastu-Kapur, S.; Ponder EL, Fonovic UP et al. (2008) Identification of proteases that regulate erythrocyte rupture by the malaria parasite *Plasmodium falciparum*, *Nature Chemical Biology*, Vol. 4, No. 3, pp. 203-210.
- Arisue, N.; Hirai M, Arai M et al. (2007) Phylogeny and evolution of the SERA multigene family in the genus *Plasmodium*, *Journal of Molecular Evolution*, Vol. 65, No. 1, pp. 82-91.
- Arisue, N.; Kawai S, Hirai M et al. (2011) Clues to evolution of the SERA multigene family in 18 *Plasmodium* species, *PLoS One*, Vol. 6, No. 3, e17775.
- Bourgon, R.; Delorenzi M, Sargeant T et al. (2004) The serine repeat antigen (SERA) gene family phylogeny in *Plasmodium*: the impact of GC content and reconciliation of gene and species trees, *Molecular Biology and Evolution*, Vol. 21, No. 11, pp. 2161-2171.
- Bzik, DJ.; Li WB, Horii T et al. (1988) Amino acid sequence of the serine-repeat antigen (SERA) of *Plasmodium falciparum* determined from cloned cDNA, *Molecular and Biochemical Parasitology*, Vol. 30, No. 3, pp. 279-288.
- Carlton, JM.; Angiuoli SV, Suh BB et al. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*, *Nature*, Vol. 419, No. 6906, pp. 512-519.
- Carlton, JM.; Adams JH, Silva JC et al. (2008) Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*, *Nature*, Vol. 455, No. 7214, pp. 757-763.
- Chulay, JD.; Lyon JA, Haynes JD et al. (1987) Monoclonal antibody characterization of *Plasmodium falciparum* antigens in immune complexes formed when schizonts rupture in the presence of immune serum. *The Journal of Immunology*, Vol. 139, No. 8, pp. 2768-2774.
- Cortés, A.; Mellombo M, Mueller I et al. (2003) Geographical structure of diversity and differences between symptomatic and asymptomatic infections for *Plasmodium falciparum* vaccine candidate AMA1. *Infection and Immunity*, Vol. 71, No. 3, pp. 1416-1426.
- Coatney, GR.; Collins WE, Warren M et al. (1971) *The Primate Malarias*, US Government Printing Office, Washington DC, USA.
- Debrabant, A.; Maes P, Delplace P et al. (1992) Intramolecular mapping of *Plasmodium falciparum* P126 proteolytic fragments by N-terminal amino acid sequencing, *Molecular and Biochemical Parasitology*, Vol. 53, No. 1-2, pp. 89-95.
- Delplace, P.; Bhatia A, Cagnard M et al. (1988) Protein p126: a parasitophorous vacuole antigen associated with the release of *Plasmodium falciparum* merozoites, *Biology of the Cell*, Vol. 64, No. 2, pp. 215-221.

- Delpace, P.; Fortier B, Tronchin G et al. (1987) Localization, biosynthesis, processing and isolation of a major 126 kDa antigen of the parasitophorous vacuole of *Plasmodium falciparum*, *Molecular and Biochemical Parasitology*, Vol. 23, No. 3, pp. 193-201.
- Duval, L.; Nerrienet E, Rousset D et al. (2009) Chimpanzee malaria parasites related to *Plasmodium ovale* in Africa, *PLoS One*, Vol. 4, No. 5, e5520.
- Felsenstein, J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods, *Methods in Enzymology*, Vol. 266, pp. 418-427.
- Fox, BA. & Bzik DJ. (1994) Analysis of stage-specific transcripts of the *Plasmodium falciparum* serine repeat antigen (SERA) gene and transcription from the SERA locus, *Molecular and Biochemical Parasitology*, Vol. 68, No. 1, pp. 133-144.
- Fox, BA.; Xing-Li P, Suzue K et al. (1997) *Plasmodium falciparum*: an epitope within a highly conserved region of the 47-kDa amino-terminal domain of the serine repeat antigen is a target of parasite-inhibitory antibodies, *Experimental Parasitology*, Vol. 85, No. 2, pp. 121-134.
- Gardner, MJ.; Tettelin H, Carucci DJ et al. (1998) Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*, *Science*, Vol. 282, No. 5391, pp. 1126-1132.
- Gardner, MJ.; Hall N, Fung E et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*, *Nature*, Vol. 419, No. 6906, pp. 498-511.
- Garzón-Ospina, D.; Cadavid LF & Patarroyo MA. (2010) Differential expansion of the merozoite surface protein (msp)-7 gene family in *Plasmodium* species under a birth-and-death model of evolution, *Molecular Phylogenetics and Evolution*, Vol. 55, No. 2, pp. 399-408.
- Gor, DO.; Li AC & Rosenthal PJ. (1998) Protective immune responses against protease-like antigens of the murine malaria parasite *Plasmodium vinckei*, *Vaccine*, Vol 16, No. 11-12, pp. 1193-1202.
- Hayakawa, T.; Culleton R, Otani H et al. (2008) Big bang in the evolution of extant malaria parasites, *Molecular Biology and Evolution*, Vol. 25, No. 10, pp. 2233-2239.
- Hayakawa, T.; Arisue N, Udonon T et al. (2009) Identification of *Plasmodium malariae*, a human malaria parasite, in imported chimpanzees, *PLoS One*, Vol. 4, No. 10, e7412.
- Hodder, AN.; Drew DR, Epa VC et al. (2003) Enzymic, phylogenetic, and structural characterization of the unusual papain-like protease domain of *Plasmodium falciparum* SERA5, *The Journal of Biological Chemistry*, Vol. 278, No. 48, pp. 48169-48177.
- Hodder, AN.; Malby RL, Clarke OB et al. (2009) Structural insights into the protease-like antigen *Plasmodium falciparum* SERA5 and its noncanonical active-site serine, *Journal of Molecular Biology*, Vol. 392, No. 1, pp. 154-165.
- Horii, T.; Shirai H, Li J et al. (2010) Evidences of protection against blood-stage infection of *Plasmodium falciparum* by the novel protein vaccine SE36, *Parasitology International*, Vol. 59, No. 3, pp. 380-386.
- Hughes, AL. & Nei M. (1989) Evolution of the major histocompatibility complex: independent origin of nonclassical class I genes in different groups of mammals, *Molecular Biology and Evolution*, Vol. 6, No. 6, pp. 559-579.
- Hughes, AL. & Nei M. (1990) Evolutionary relationships of class II major-histocompatibility-complex genes in mammals, *Molecular Biology and Evolution*, Vol. 7, No. 6, pp. 491-514.
- Janssen, CS.; Phillips RS, Turner CM et al. (2004) *Plasmodium* interspersed repeats: the major multigene superfamily of malaria parasites, *Nucleic Acids Research*, Vol. 32, No. 19, pp. 5712-5720.

- Kiefer, MC.; Crawford KA, Boley LJ et al. (1996) Identification and cloning of a locus of serine repeat antigen (sera)-related genes from *Plasmodium vivax*, *Molecular and Biochemical Parasitology*, Vol. 78, No. 1-2, pp. 55-65.
- Knapp, B.; Hundt E, Nau U et al. (1989) Molecular cloning, genomic structure and localization of a blood stage antigen of *Plasmodium falciparum* characterized by a serine stretch. *Molecular and Biochemical Parasitology*, Vol. 32, No. 1, pp. 73-83.
- Knapp, B.; Nau U, Hundt E et al. (1991) A new blood stage antigen of *Plasmodium falciparum* highly homologous to the serine-stretch protein SERP, *Molecular and Biochemical Parasitology*, Vol. 44, No. 1, pp. 1-13.
- Levin, ND. (1988) *The protozoan phylum Apicomplexa-volume II*, CRC Press, ISBN 0-8493-4654-1, Florida, USA.
- Li, J.; Mitamura T, Fox BA et al. (2002a) Differential localization of processed fragments of *Plasmodium falciparum* serine repeat antigen and further processing of its N-terminal 47 kDa fragment, *Parasitology International*, Vol. 51, No. 4, pp. 343-352.
- Li J.; Matsuoka H, Mitamura T et al. (2002b) Characterization of proteases involved in the processing of *Plasmodium falciparum* serine repeat antigen (SERA), *Molecular and Biochemical Parasitology*, Vol. 120, No. 2, pp. 177-186.
- Liu, Q.; Ferreira MU, Ndawi BT et al. (2000) Sequence diversity of serine repeat antigen gene exon II of *Plasmodium falciparum* in worldwide collected wild isolates. *Southeast Asian Journal of Tropical Medicine and Public Health*, Vol. 31, No. 4, pp. 808-817.
- McBride, JS.; Newbold CI & Anand, R. (1985) Polymorphism of a high molecular weight schizont antigen of the human malaria parasite *Plasmodium falciparum*, *The Journal of Experimental Medicine*, Vol. 161, No. 1, pp. 160-180.
- McCoubrie, JE.; Miller SK, Sargeant T et al. (2007) Evidence for a common role for the serine-type *Plasmodium falciparum* serine repeat antigen proteases: Implications for vaccine and drug design, *Infection and Immunity*, Vol. 75, No. 12, pp. 5565-5574.
- Miller, SK.; Good RT, Drew DR et al. (2002) A subset of *Plasmodium falciparum* SERA genes are expressed and appear to play an important role in the erythrocytic cycle, *The Journal of Biological Chemistry*, Vol. 277, No. 49, pp. 47524-47532.
- Morimatsu, K.; Morikawa T, Tanabe K et al. (1997) Sequence diversity in the amino-terminal 47 kDa fragment of the *Plasmodium falciparum* serine repeat antigen. *Molecular and Biochemical Parasitology*, Vol. 86, No. 2, pp. 249-254.
- Nei, M. & Hughes AL. (1992) Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. In: *11th Histocompatibility Workshop and Conference*, Tsuji, K., Aizawa, M., Sasazuki, T., editors. Vol. 2, pp. 27-38, Oxford University Press, ISBN 0-19-262217-X, Oxford, UK.
- Nei, M.; Gu, X. & Sitnikova T. (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 94, No. 15, pp. 7799-7806.
- Nei, M. & Rooney AP. (2005) Concerted and birth-and death evolution of multigene families, *Annual Review of Genetics*, Vol. 39, pp. 121-152.
- Nishimoto, Y.; Arisue N, Kawai S et al. (2008) Evolution and phylogeny of the heterogeneous cytosolic SSU rRNA genes in the genus *Plasmodium*, *Molecular Phylogenetics and Evolution*, Vol. 47, No. 1, pp. 45-53.
- Okech, BA.; Nalunkuma A, Okello D et al. (2001) Natural human immunoglobulin G subclass responses to *Plasmodium falciparum* serine repeat antigen in Uganda, *The American Journal of Tropical Medicine and Hygiene*, Vol. 65, No. 6, pp. 912-917.

- Okech, B.; Mujuzi G, Ogwal A et al. (2006) High titers of IgG antibodies against *Plasmodium falciparum* serine repeat antigen 5 (SERA5) are associated with protection against severe malaria in Ugandan children, *The American Journal of Tropical Medicine and Hygiene*, Vol. 74, No. 2, pp. 191-197.
- Okitsu, SL.; Boato F, Mueller MS et al. (2007) Antibodies elicited by a virosomally formulated *Plasmodium falciparum* serine repeat antigen-5 derived peptide detect the processed 47 kDa fragment both in sporozoites and merozoites, *Peptides*, Vol. 28, No. 10, pp. 2051-2060.
- Ota, T. & Nei M. (1994) Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. *Molecular Biology and Evolution*, Vol. 11, No. 3, pp. 469-482.
- Pain, A.; Bohme U, Berry AE et al. (2008) The genome of the simian and human malaria parasite *Plasmodium knowlesi*, *Nature*, Vol. 455, No. 7214, pp. 799-803.
- Palacpac, NM.; Leung BW, Arisue N et al. (2006) *Plasmodium vivax* serine repeat antigen (SERA) multigene family exhibits similar expression patterns in independent infections, *Molecular and Biochemical Parasitology*, Vol. 150, No. 2, pp. 353-358.
- Perrin, LH.; Merkli B, Loche M et al., (1984) Antimalarial immunity in Saimiri monkeys. Immunization with surface components of asexual blood stages, *The Journal of Experimental Medicine*, Vol. 160, No. 2, pp. 441-451.
- Polley, SD.; Chokejindachai W & Conway DJ. (2003) Allele frequency-based analyses robustly map sequence sites under balancing selection in a malaria vaccine candidate antigen, *Genetics*, Vol. 165, No. 2, pp. 555-561.
- Putrianti, ED.; Schmidt-Christensen A, Arnold I et al. (2010) The *Plasmodium* serine-type SERA proteases display distinct expression patterns and non-essential in vivo roles during life cycle progression of the malaria parasite, *Cellular Microbiology*, Vol. 12, No. 6, pp. 725-739.
- Rooney, AP. (2004) Mechanism underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in apicomplexans, *Molecular Biology and Evolution*, Vol. 21, No. 9, pp. 1704-1711.
- Schmidt-Christensen, A.; Sturm A, Horstmann S et al. (2008) Expression and processing of *Plasmodium berghei* SERA3 during liver stages, *Cellular Microbiology*, Vol. 10, No. 8, pp. 1723-1734.
- WHO (2010) *World Malaria Report 2010*, WHO Press, ISBN 978 92 4 156410 6, Geneva, Switzerland.
- Yeoh, S.; O'Donnell RA, Koussis K et al. (2007) Subcellular discharge of a serine protease mediates release of invasive malaria parasites from host erythrocytes, *Cell*, Vol. 131, No. 6, pp. 1072-1083.
- Zhang J.; Dyer KD. & Rosenberg HF. (2000) Evolution of the rodent eosinophil-associated RNase gene family by rapid gene sorting and positive selection, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 97, No. 9, pp. 4701-4706.

# Molecular Evolution of Juvenile Hormone Signaling

Aaron A. Baumann and Thomas G. Wilson  
*The Ohio State University*  
*United States of America*

## 1. Introduction

Insect development proceeds through a series of discrete developmental stages called instars. During hexapod evolution, the development of complete metamorphosis introduced a novel mechanism for separating feeding and reproductive stages (Truman & Riddiford, 2002), facilitating the tremendous evolutionary success of holometabolous insects. In contrast to hemimetabolous insects, which progress through a series of instars that appear as smaller iterations of the adult form, holometabolous insects proceed from egg to adult through a progression of isomorphic larval instars and a pupal transitory stage. In each case, the physical boundary for growth during an instar is established by a chitinous exoskeleton, which must be periodically shed. This molting process is under the control of two counteracting hormones.

Toward the end of an instar, a pulse of the insect molting hormone, 20-hydroxyecdysone (20E) initiates a transcriptional cascade that carries the molt to a subsequent instar. However, it is the interaction of 20E and the sesquiterpenoid juvenile hormone (JH) that governs the developmental outcome of each molt. During larval development, an elevated JH titer and 20E directs the sequential progression through larval development until the final larval instar, when the JH titer substantially declines. The removal of circulating JH facilitates a 20E-directed developmental switch that initiates the metamorphic molt. Thus, it was proposed that JH can modulate 20E activity, maintaining the status quo during pre-adult development.

### 1.1 JH and JHAs: Insecticidal use of hormone agonists

Since the first chemical analysis resolved the sesquiterpenoid structure of endogenous JH (Röller *et al.*, 1967), several homologs have been identified, each bearing opposing, terminal epoxide and methyl ester functions. Variation in the degree and identity of alkyl group substitution at C3, C7, and C11 along the carbon skeleton defines the homologs. The evolutionary importance of multiple JH homologs is unclear. JH 0, I, II, and III have all been isolated from lepidopteran insects, whereas JH III, the presumed evolutionary precursor to the higher homologs, is found in all insects. JH bisepoxide (JHB3) has been identified as a product of the corpus allatum (CA) in higher Diptera including *Drosophila melanogaster* and *Sarcophaga bullata* (Richard *et al.*, 1989; Bylemans *et al.*, 1998). Nearly identical in structure to JH III, JHB3 is distinguished by an additional epoxide group spanning C6-C7.

The major JHs and some juvenile hormone analogs (JHAs) are presented in Figure 1.

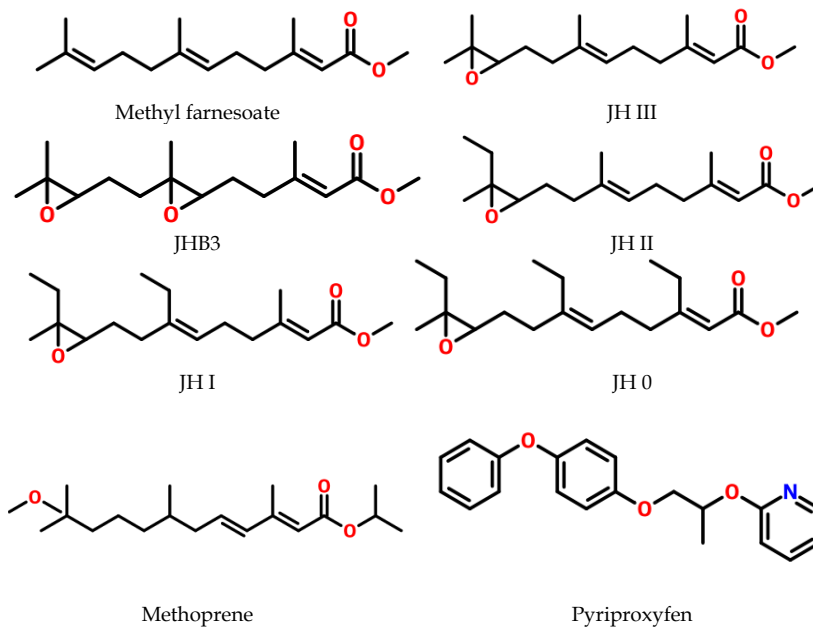


Fig. 1. Structures of endogenous JH molecules and two synthetic JHAs, methoprene and pyriproxyfen.

The physiology and chemistry of JH prompted intense research into the synthesis and commercial-scale production of JH analogs, or juvenoids, for agricultural use. The allure of these compounds was at least twofold. First, juvenoids exhibit extremely low non-target (in particular, mammalian) toxicity. Second, it was originally thought that insect resistance to JHAs would be unlikely, since an insect was not likely to become refractory to an endogenous hormone (Williams, 1967). Methoprene, a juvenoid structurally similar to endogenous JH, has enjoyed success in the management of larval mosquito populations. However, JHAs need not mimic the chemical structure of endogenous JH, as exemplified by the pyridine-based pyriproxyfen, whose activity exceeds JH by two orders of magnitude in dipteran white puparial and larval assays (Riddiford and Ashburner, 1991).

Exogenous JH exposure can elicit classic antimetamorphic activity in both Lepidoptera and Coleoptera (Srivastava & Srivastava, 1983; Konopova & Jindra, 2007), extending larval development through one or more supernumerary instars. Also in these insects, exposure to exogenous JH or to its chemical analogs (JHA) can result in the deposition of a second pupal cuticle (Zhou & Riddiford, 2002). Thus, in Lepidoptera and Coleoptera, JH exposure at an inappropriate time inhibits 20E-directed developmental progression.

In Diptera, treatment with exogenous JH produces dose-dependent lethality at the pharate adult stage. All adult structures arise from imaginal discs in flies, and these discs are insensitive to JH during development, unlike Lepidoptera and Coleoptera, in which the

polymorphic larval epidermis gives rise to pupal and adult structures. When flies are challenged with JHAs, the adult structures that differentiate from imaginal discs remain unaffected (Postlethwait, 1974). In *D. melanogaster*, only the abdominal histoblasts are JH sensitive; diagnostic (sublethal) doses of methoprene disrupt abdominal bristle formation in female flies (Madhavan, 1973).

## 2. Molecular mechanism of JH signal transduction

The molecular events underlying 20E signaling are relatively well understood. Ecdysone released from the prothoracic glands is converted to its active metabolite 20E in target tissues, where it regulates transcription through a heterodimeric receptor complex comprised of *Ecdysone receptor* (EcR) and *Ultraspiracle* (USP) proteins. When bound with 20E, ECR-USP recognizes and binds ecdysone response elements located in the promoter region of target genes, inducing transcription of a hierarchical network of early and late genes. The early genes either repress their own expression or induce expression of late genes (Ashburner *et al.*, 1974). In this manner, the expression of genes involved in the 20E transcriptional cascade is tightly controlled. In contrast, the nature of JH signal transduction has been difficult to elucidate, largely due to the enigmatic nature of the JH receptor. A body of ever-increasing experimental evidence strongly supports the product of the *Methoprene tolerant* (*Met*) gene as the prime candidate for a JH receptor component (Wilson & Fabian, 1986; Konopova & Jindra, 2007; Yang *et al.*, 2011).

*Met* was originally discovered by screening progeny of ethyl methanesulfonate (EMS)-mutagenized *D. melanogaster* for resistance to methoprene (Wilson & Fabian, 1986). *Met* mutants show dramatically enhanced (~100 fold) resistance to both the toxicity and morphogenetic defects caused by methoprene exposure, but not to other classes of insecticides (Wilson & Fabian, 1986). Such resistance is not restricted to compounds with high structural similarity to JH; *Met* mutants are also resistant to the more potent, structurally distinct JHA pyriproxyfen (Riddiford & Ashburner, 1991).

Cloning and sequence analysis identified *Met* as a member of the basic Helix-Loop-Helix *Period Ahr Sim* (bHLH PAS) family of transcriptional regulators (Ashok *et al.*, 1998). PAS proteins function as dimers in a diverse array of functions in development, xenobiotic binding, and detection of environmental signals (Crews, 1993). Both the bHLH domain and the PAS repeats (PAS A and B) facilitate dimerization between PAS proteins (Huang, *et al.*, 1993). Additionally, the PAS domains function in small molecule ligand binding and target gene specificity. Each dimerization partner recognizes and binds one half of a palindromic E-box consensus sequence CANNTG in the promoter region of target genes via the stretch of basic residues immediately N-terminal to the HLH motif. Examples of PAS proteins with ligand binding activity include the bacterial *photoreactive yellow protein* (PYP), and the vertebrate *aryl hydrocarbon receptor* (Ahr).

Genetic and biochemical data show that MET binds JH with nanomolar affinity (Shemshedini & Wilson, 1990) and that MET product is present in the nuclei of several known JH target tissues, including ovary, MAG, and larval fat body (Pursley *et al.*, 2000). In addition, MET can drive the expression of a reporter gene in a JH-sensitive manner (Miura *et al.*, 2005). All of the above data satisfy criteria for a hormone receptor.

Analysis of the *Met*<sup>27</sup> null allele provided the first demonstration of insecticide resistance due to the absence of a target macromolecule (Wilson & Ashok, 1998). Even though *Met*<sup>27</sup> flies are viable, *Met* deficiency carries reproductive consequences, namely substantially

reduced oogenesis (~20% compared to *Met*<sup>+</sup>), consistent with a role for JH in this physiology. However, since absence of a JH receptor is expected to preclude normal development, the viability of *Met*<sup>27</sup> flies challenged the notion of *Met* as a *bona fide* JH receptor. Some evidence supports alternative mechanism(s) of JH signaling (see Flatt *et al.*, 2008; Riddiford *et al.*, 2010). In this chapter, we review data that support the notion of *germ cell expressed (gce)*, the paralog of *Met* in higher Diptera, as conferring at least partial functional redundancy.

### 3. *Met* homologs across holometabola

Reports of methoprene resistance in mosquito populations (Dame *et al.*, 1998; Cornel *et al.*, 2000; Cornel *et al.*, 2002) led us to investigate the *Met* orthologs of three mosquito species: *Aedes aegypti*, *Culex pipiens*, and *Anopheles gambiae*. Using a combination of degenerate RT-PCR and genomic database mining, we isolated a single *Met* homolog from each of these mosquitoes. Sequence analysis of these genes showed that they share high identity with both *Met* and the closely related *gce* from *Drosophila*, as expected. However, a comparison of the genomic structures among *DmMet*, *Dmgce*, and the three putative mosquito *Met* genes revealed higher structural conservation between each mosquito *Met* and *Dmgce*. Importantly, the intron number of these genes is more consistent with that of *Dmgce* than *DmMet* (from six to nine, *versus* one in *DmMet*). Furthermore, several introns in each mosquito gene are positionally conserved with those in *Dmgce*. This led to our proposal that the *Met* gene of higher Diptera originated via retrotransposition of a basal, *gce*-like gene of lower Diptera (Wang *et al.*, 2007).

Retrotransposition, or retroposition is a mechanism of gene duplication that proceeds through an mRNA intermediate. Following post-transcriptional splicing, the parental message is reintegrated into the genome. Ultimately, for the duplicate copy to escape the fate of becoming a pseudogene, it must reintegrate with associated regulatory elements intact or incorporate into a suitable transcriptional environment elsewhere in the genome. Following duplication, the increase in copy number of the parental gene affords a relaxation of selective constraint, facilitating functional divergence. This may manifest as subfunctionalization, in which a modification of the parental function evolves, or neofunctionalization, which refers to attainment of a novel function (MacCarthy & Bergman, 2007). *DmMet* retains a strong diagnostic feature of retroposition: a paucity of introns relative to *gce*, which is consistent with splicing and genomic reintegration of an ancestral *gce*-like transcript.

A conserved *gce*-like gene appears to be conserved across holometabolan genomes, including the red flour beetle, *Tribolium castaneum*, and the honeybee, *Apis mellifera*. An independent gene duplication within the Lepidoptera has given rise to two *Met*-like proteins, presently called Methoprene tolerant proteins I and II, whose functions are currently under investigation (i.e. Li *et al.*, 2010). Despite a demonstrated sequence conservation favoring the *Met*-like genes of more primitive Holometabola as ancestral to *gce*, we will continue to refer to these genes as *Met*-like in this text.

#### 3.1 *Met* and *gce* within the genus *Drosophila*

When the genomes of 12 representative *Drosophila* species became available (Ashburner, 2007), we chose to examine the molecular evolution of *Met* and *gce* within this genus of flies. Both paralogs are conserved in each species, indicating that the origin of *Met* predates that



of the genus *Drosophila*, some 63 million years ago (Tamura *et al.*, 2004). The architecture of these genes is generally conserved in each species, with a few notable exceptions. A single conserved intron is present in *Met* in the PAS B domain of 11 species. In addition to this conserved intron, independent intron gains have occurred in the lineages leading to *D. simulans* and *D. willistoni*. A single *Met* ortholog exists in each *Drosophila* genome examined, but *D. persimilis* harbors two separate, consecutive loci on the X chromosome, currently called GL13106 and GL13107, that align to distinct regions of *DmMet*. The 5' putative gene GL13106 contains a complete PAS A domain followed by a severely truncated PAS B domain. We performed RT-PCR across these two genes and failed to obtain a single PCR product, suggesting that GL13106 and GL13107 indeed code for two distinct open reading frames. Eleven of the 12 representative *gce* orthologs contain at least six conserved introns, with independent intron gains evident in the lineages leading to *D. melanogaster*, *D. pseudoobscura*, and *D. mojavensis*, whereas a substantial deletion in *D. persimilis gce* has eliminated the central portion of this gene, including the PAS repeats.

In addition to the bHLH, PAS, and PAC domains, putative transactivation domains (TAD) are evident in *Met* and *gce* orthologs. TADs are glutamine and/or aspartic acid-rich motifs whose amino acid sequences are broadly defined and generally reside in the C-terminal region of PAS proteins (Ramadoss & Perdew, 2005). *Met* homologs show Q- and D-rich motifs between the PAS B and PAC domains, while alignments of *gce* homologs indicate a D-rich region C-terminal to the PAC domain. Miura *et al.* (2005) suggest the presence of a C-terminal TAD in recombinant MET protein, but this region has yet to be functionally defined.

Using *DmMet* and *Dmgce* as query sequences, we conducted homology searches under tBLASTx criteria (translated nucleotide query to search a translated nucleotide database) against the publicly available EST library of *Glossina morsitans*, the tsetse fly. Our search recovered several clones, which were imported into the Sequencher program to produce two independent contigs. These composite nucleotide sequences were used to infer a gene tree with other holometabolan *Met* and *gce* orthologs, including those of two representative *Drosophila* species (Figure 2). This preliminary analysis reveals the presence of distinct *Met* and *gce* orthologs in the *G. morsitans* genome, indicating that the origin of *Met* predates the divergence of the Aschiza and Schizophora. These two taxonomic groups, which are estimated to have diverged more than 85 million years ago (Bertone & Wiegmann, 2009), reside within the brachyceran infraorder Muscomorpha.

### 3.2 Evidence for differential selective constraint imposed on *Met* and *gce*

Based on an *a priori* hypothesis that *Met* and *gce* were subject to differential post-duplication selective constraint, we performed analyses of nonsynonymous-to-synonymous (dN/dS) substitution ratios on codon alignments of these *Drosophila* paralogs. Datasets were analyzed using the DataMonkey tool (Kosakovsky-Pond & Frost, 2005), a web-based implementation of the HyPhy package (Kosakovsky Pond *et al.*, 2005). dN/dS analyses can be used to infer the relative selective pressure along entire coding sequences or in a site-specific manner. A substantially depressed dN/dS ratio (i.e. zero or close to zero) implies purifying (negative) selection. That is, nonsynonymous changes are stringently selected against. In contrast, when dN/dS is nearly one, neutral evolution is inferred. A dN/dS value far in excess of one implies positive selection, or adaptive evolution. In this case, nonsynonymous substitutions confer a selective advantage.

The results of our dN/dS analyses showed dramatic dissimilarity in the relative selective pressures that have shaped the coding sequences of *Met* and *gce*. In the case of *Met*, dN/dS was generally suppressed along the entirety of the coding sequence, indicating strong selection against nonsynonymous codon substitution. This is perhaps surprising, since MET deficiency has no effect on viability (Wilson & Ashok, 1998). Possibly, mutations that alter amino acid identities are selected against in *Met* due its involvement in reproduction. In the absence of methoprene selection, *Met* mutants are quickly out-competed by wild type flies despite the seemingly slight fitness cost of *Met* loss (Minkhoff III & Wilson, 1992). In contrast, dN/dS values close to one dominate the C-terminal half of *gce*, indicating a substantial relaxation of selective constraint in this region. The N-terminal region of this gene, containing the canonical bHLH and PAS functional domains, shows a strongly depressed dN/dS. Based on functional data from other PAS proteins, this region is assumed to harbor DNA and ligand binding activity, whereas the C-terminal region contains putative TADs. C-terminal degeneracy was shown to confer differential target gene specificity between the *Ahr* homologs of mice and humans (Ramadoss & Perdeu, 2005; Flaveny *et al.*, 2010). Similarly, the disparate selective constraints evident in the C-terminal regions of *Met* and *gce* may partially define these genes' functions.

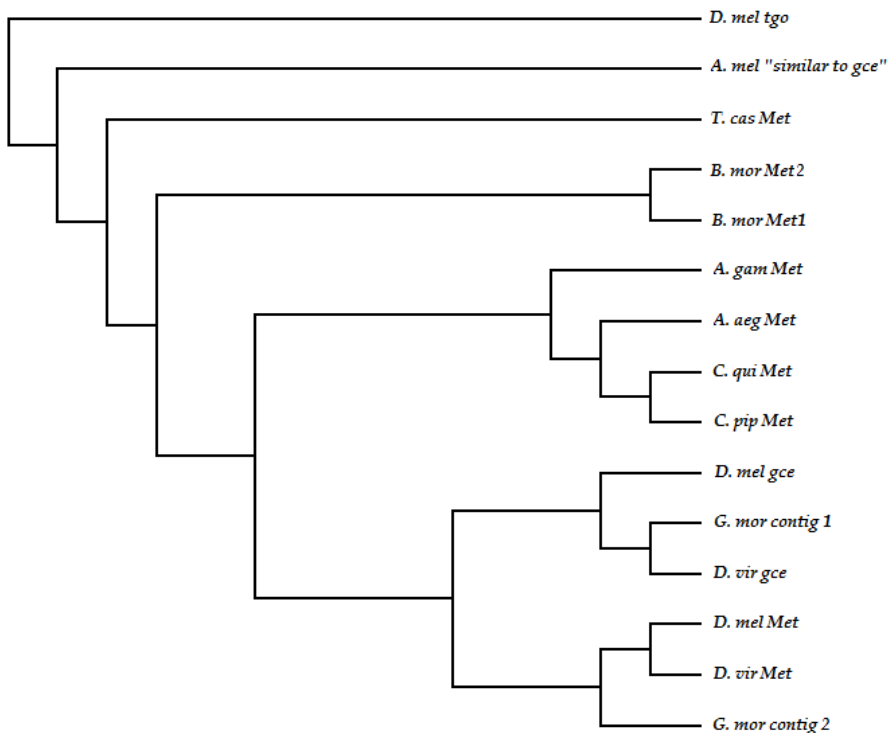


Fig. 2. A gene tree of some holometabolous *Met*-like genes, showing placement of two distinct *G. morsitans* sequences as putative *Met* and *gce* orthologs. *D. melanogaster* Tango (*tgo*), the homolog of the vertebrate Aryl hydrocarbon receptor (*Ahr*), is used as an outgroup sequence.

#### 4. Toward a functional definition of *Dmgce*

A functional characterization of *gce*, named for its expression in a subset of embryonic germ cells (Moore *et al.*, 2000), is in its infancy. Column pulldown assays showed MET, in addition to forming homodimers, forms heterodimers with GCE, and addition of JH or either of two JHAs significantly impaired these interactions (Godlewski *et al.*, 2006). It is unknown whether GCE forms homodimers, like MET, or whether GCE can bind JH or its analogs. *GAL4/UAS*-driven (Brand & Perrimon, 1993) overexpression of *Met*<sup>+</sup> from *actin* or *tubulin* promoters results in larval lethality in the absence of methoprene (Barry *et al.*, 2008), perhaps by upsetting the stoichiometry of MET and GCE dimers, favoring MET homodimerization at inappropriate times or in inappropriate tissues. Recently, JH was shown to inhibit MET and GCE in *D. melanogaster* by preventing caspase-driven programmed cell death (PCD) and histolysis of the larval fat body. DRONC and DRICE, evolutionarily conserved caspase genes involved in this physiology at the onset of metamorphosis, were shown to be downregulated in *Met* and *gce* deficient flies (Liu *et al.*, 2009). Similarly, methoprene interferes with caspase-driven midgut remodeling in *A. aegypti* (Nishiura *et al.*, 2003; Wu *et al.*, 2006) and *T. castaneum* (Parthasarathy *et al.*, 2008; Parthasarathy *et al.*, 2009), showing that this mechanism of JH action is evolutionarily conserved. It is noteworthy that recombinant MET can repress reporter gene expression in the absence of JH (presumably, MET forms homodimers in this system; Miura *et al.*, 2005); transcriptional repression has previously been reported in other PAS proteins (Dolwick *et al.*, 1993). Therefore, the JH-dependent, stage-specific formation of alternative MET/GCE dimers may have unique regulatory consequences on distinct suites of target genes.

##### 4.1 *Dmgce* substitution for *DmMet*

To evaluate the notion that *gce* might confer viability to *Met* null flies, we manipulated *gce* expression using a binary *UAS/GAL4* system to drive either a *gce* cDNA or an RNAi construct designed to target *gce* transcript. We carried these experiments out in a variety of genotypic contexts in order to examine the effect of *gce* transcript abundance on several methoprene conditional and non-conditional phenotypes (Baumann *et al.*, 2010b). First, we explored the effect of *gce* over- and under-expression on a *Met*-specific non-conditional phenotype that manifests as a variable number of grossly malformed posterior facets of the compound eye (Figure 3). This phenotype is visible in *Met*<sup>27</sup> and *Met*<sup>w3</sup> flies, and is enhanced in the latter genotype. In our experiments, we found that *gce* overexpression in a *Met*<sup>w3</sup> genetic background can rescue the *Met*-specific eye phenotype, suggesting functional overlap of *gce* and *Met*. Notably, when *gce* was overexpressed in a *Met*<sup>27</sup> background from the *GawB*;*dan*[*AC116*] promoter, targeting transgene expression to the compound eye, the eye phenotype was completely rescued (Baumann *et al.*, 2010b).

The *Met*<sup>27</sup> phenotype mimics a set of defects resulting from genetic ablation of the JH-producing corpus allatum (CAX), including a heterochronic shift in EcR-B1 expression in the optic lobe (Riddiford *et al.*, 2010). Exogenous JH application rescues the entire suite of defects in CAX prepupae, while JH provision to *Met*<sup>27</sup> flies rescues only a subset of these defects, suggesting an alternate mechanism of JH signal transduction (Riddiford *et al.*, 2010). Based on our findings that *gce* can substitute for *Met* in the compound eye, further study of GCE involvement in eye development may provide a link between these phenomena. For instance, GCE may partially substitute for MET as a ligand binder to mediate JH signaling when this hormone is supplied in excess.

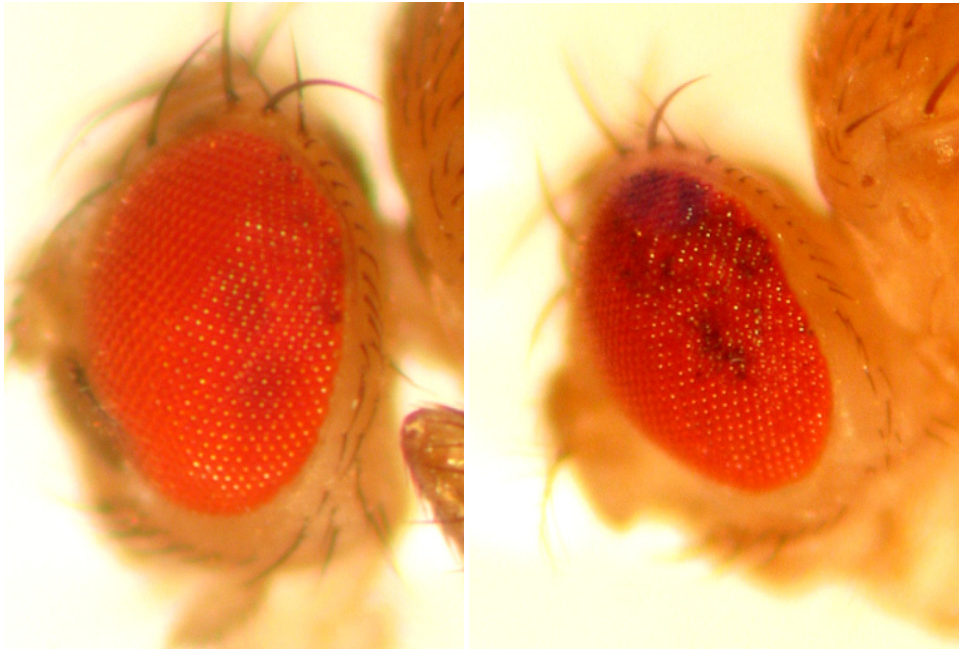


Fig. 3. Left: malformed facets in the posterior compound eye of *Met<sup>w3</sup>* flies appear dark under light microscopy. Right: EMS-induced production of an unidentified enhancer gene dramatically intensifies the *Met<sup>w3</sup>* phenotype (T.G.W., unpublished).

We also explored the effect of *gce* overexpression on several methoprene-conditional phenotypes. Overexpressed *gce* rescued both the diagnostic malrotation of male genitalia and sensitivity to the toxic effects of methoprene exposure. Sublethal doses of methoprene can induce malrotation of the male genital disc in *D. melanogaster*, resulting in terminalia that are improperly oriented for copulation (Bouchard & Wilson, 1987). *Met<sup>27</sup>* males are resistant to this phenotype. We found that global *gce* overexpression in a *Met<sup>27</sup>* background rescues blockage of the malrotation phenotype in *Met<sup>27</sup>; UAS-gce/tubulin-GAL4* flies. When these flies were exposed to methoprene, we observed malrotation close to levels seen in *Met<sup>+</sup>* flies (Baumann *et al.*, 2010b).

*Met* and *gce* are generally co-expressed in JH target tissues, but we detected insignificant amounts of *gce* transcript in late third instar larval fat body. When *gce* was expressed from a construct targeting expression to this tissue, partial rescue of JH-induced pupal lethality was achieved, perhaps as a result of supplying *gce* to a tissue in which its expression is normally depressed at this time in development. *gce* expression in the larval fat body was unable to rescue either the eye phenotype or to prevent methoprene-induced malrotation of the male genitalia, indicating that *gce* substitution for *Met* is tissue specific (Baumann *et al.*, 2010b).

#### 4.2 Functional partitioning of *DmMet* and *Dmgce* in *D. melanogaster* reproduction

Following metamorphosis, the interaction of 20E and JH is crucial in insect reproduction. JH was first isolated in large quantities from the MAG of *Hyalophora cecropia* (Williams,

1956), suggesting a role in male reproductive biology. In *D. melanogaster*, JH controls MAG protein accumulation (Yamamoto *et al.*, 1988) and male *apterous* (*ap*) mutants court females less vigorously than wild-type flies (Tompkins, 1990). In females, the activity of these counteracting hormones is critical for ovarian development and oocyte maturation. Development of the *D. melanogaster* oocyte is under the control of JH through previtellogenic stages 8-9. Female *D. melanogaster apterous*<sup>4</sup> mutants are sterile owing to reduced levels of JH synthesis (Bownes, 1989); provision of exogenous JH rescues vitellogenic oocyte development in *ap* females (Postlethwait & Weiser, 1973). In *A. aegypti*, JH also controls previtellogenic ovarian development (Clements, 1992). In this case, JH signaling is necessary to promote 20E competence in the fat body, the site of post-blood meal vitellogenin synthesis. In contrast, vitellogenesis is retarded by JH treatment in the gypsy moth, *Lymantria dyspar* (Davis *et al.*, 1990). Thus, there is variation in hormonal control in insects.

GCE clearly compensates for MET deficiency in preadult development (Baumann *et al.*, 2010b). In our experiments, over-expressed *gce* failed to rescue both the documented behavior of reduced courtship in *Met*<sup>27</sup>; *UAS-gce/tubulin-GAL4* males and the reduction in oocyte development and oviposition in these females. Therefore, it appears that excess *gce* cannot compensate for *Met*-induced reduction of reproductive capacity. This result suggests that the functional roles for MET and GCE are incompletely partitioned between preadult development and reproduction in adults.

In *A. aegypti*, AaMet regulates the transcription of several JH target genes in newly eclosed, previtellogenic adult females (Zhu *et al.*, 2010). Presumably, the MET-like gene product in lower Diptera serves an analogous function both MET and GCE in JH signaling, but through the action of a single gene. This is perhaps accomplished by virtue of its modular architecture of *DmMet*- and *Dmgce*-specific domains. Higher sequence identity exists between the bHLH and PAS B of *Dmgce* and more primitive holometabolous *Met*-like genes, while the PAS A and PAC domains share higher sequence identity with *DmMet*. These domains may confer a discriminating *Met*-like function that may partially underlie the functional divergence of *Met* and *gce* in higher Diptera.

### 4.3 *Dmgce* is a vital gene

Overexpression studies demonstrated that *gce* can substitute for *Met* in a tissue specific manner to rescue several preadult *Met* mutant phenotypes. Hence, our results empirically support the notion of functional redundancy between *Met* and its paralog *gce*. To further explore the relationship between *Met* and *gce* in JH signaling, we carried out underexpression studies in *Met*<sup>+</sup> and *Met* mutant backgrounds by driving the expression of a *gce* RNAi construct.

First, we examined the consequence of *gce* deficiency in a *Met* mutant background under the justification that, if *gce* is responsible for *Met*<sup>27</sup> viability, then concomitant reduction of *Met* and *gce* could result in lethality. Interestingly, *Met*<sup>27</sup>; *UAS-gce-dsRNA/tubulin-GAL4* flies died as early pupae (0-2 days), whereas expression of the dsRNA construct from an *actin-GAL4* promoter caused lethality in the pharate adult stage. Next, we assessed the effects of *gce* reduction in *Met*<sup>+</sup> flies. Surprisingly, *Met*<sup>+</sup>; *UAS-gce-dsRNA/tubulin-GAL4* flies died as pharate adults, indicating that even in the presence of functional MET, *gce* is a vital gene. Driving the transgene from an *actin-GAL4* promoter allowed some degree of adult survival, but these adults were clearly affected by insufficient *gce*, dying within two to three days.

Differential intensity of transgene expression from *actin* and *tubulin* promoters was previously reported in our lab (Barry *et al.*, 2008).

*gce* underexpression had no observable effect on embryonic development, a stage during which no role for JH has been demonstrated. We have shown that *gce* transcription begins after about eight hours in early embryos, in contrast to *Met*, which is supplied as a maternal message (Baumann *et al.*, 2010a). The importance of such divergence in temporal expression profiles is unclear.

## 5. Evolutionary conservation of JH signaling mechanisms

Numerous JH target genes have been identified throughout Holometabola. Importantly, many of these genes are known components of the early 20E response. Table 1 lists some representative JH-inducible genes.

Symbol	Gene name	Molecular function	Reference
<i>Jhl-1</i>	JH inducible protein 1	Endoribonuclease	Dubrovsky <i>et al.</i> , 2000
<i>Jhl-26</i>	JH inducible protein 26	Unknown	
<i>Br</i>	Broad-Complex (BR-C)	BTB POZ zinc finger transcription factor	Zhou <i>et al.</i> , 1998; Zhou & Riddiford, 2002
<i>mnd</i>	Minidisks	Amino acid transmembrane transporter	Dubrovsky <i>et al.</i> , 2002
<i>Jhl-21</i>	JH inducible protein 21	Amino acid transmembrane transporter	
<i>JHE</i>	JH esterase	JH-specific esterase	Kethidi <i>et al.</i> , 2005
<i>E75A</i>	Ecdysone-induced protein 75B	Heme binding	Dubrovsky <i>et al.</i> , 2004
<i>E74B</i>	Ecdysone-induced protein 74EF	RNA polymerase II transcription factor activity	Beckstead <i>et al.</i> , 2007
<i>pepck</i>	Phosphoenolpyruvate carboxykinase	Phosphoenolpyruvate carboxykinase (GTP) activity	
CG14949	CG14949	Unknown	

Table 1. Representative JH-inducible genes. Many of these genes have evolutionarily conserved roles in JH signaling in holometabolous insects. In addition, several are known components of the 20E transcriptional cascade.

The majority of the work done in our lab has been carried out on *D. melanogaster*, in which *DmMet* clearly plays a role in JH signaling: its absence both interferes with methoprene toxicity (Wilson & Fabian, 1986) and hinders JH-driven reproductive physiology (Wilson, 1992; Wilson *et al.*, 2003). However, *Met* involvement in metamorphosis has been difficult to demonstrate in *Drosophila* (Riddiford, 2008). As previously stated, JH exposure has no effect on dipteran entry into metamorphosis, unlike other insects (Williams, 1961; Zhou & Riddiford, 2002). In recent years, researchers have turned to the model coleopteran, *T.*

*castaneum*. These beetles are both amenable to genetic manipulation and gene knockdown owing to the dramatic effects of systemic RNAi, and the larvae of this species are very sensitive to JH, unlike *D. melanogaster* larvae. Exposure to JH or a number of its chemical analogs precipitates supernumerary larval instars, similar to the effects of JH on the model lepidopteran, *Manduca sexta* (Parthasarathy & Palli, 2009). *T. castaneum*, like mosquitoes, has a single *Met-like* gene. In their seminal paper, Konopova and Jindra (2007) demonstrated that RNAi-mediated knockdown of *TcMet* results not only in a methoprene resistance phenotype, but also in the precocious metamorphosis of early instar larvae. A long sought-after result, the genetic reduction of *TcMet* provided the phenotype frustratingly absent in *D. melanogaster*: metamorphic disruption. Reproductive roles for *TcMet* have also been shown; *TcMet* knockdown results in a substantial decrease in vitellogenin transcription, (Parthasarathy, *et al.*, 2010) consistent with *Met* deficiency in *D. melanogaster* females (Wilson & Ashok, 1998). These results demonstrate that the single *Met-like* genes in primitive Holometabola function in both development (metamorphosis) and reproduction. Further functional characterization of *TcMet* (and the single *Met-like* gene of lower Diptera) could lead to a better understanding of how *DmMet* has apparently co-opted reproductive functional roles from a *gce*-like ancestor in higher Diptera

### 5.1 JH regulation of the E-20 transcriptional cascade

The molecular networks that link JH and 20E signaling pathways form the foundation of multiple aspects of insect physiology, as evidenced by the criticality of both hormones in development, reproduction, and diapause (Zhou & Riddiford, 2002; Soller *et al.*, 1999; Denlinger, 1985). *Broad Complex* (*Broad* or *BR-C*) is an early gene in the 20E cascade that encodes a family of alternatively spliced zinc finger transcription factors (four in *D. melanogaster*, Z1-Z4) fused to a common core protein. Certain *Broad* alleles phenocopy the morphogenetic defects incurred by methoprene exposure in *D. melanogaster*. Wilson *et al* (2006) showed phenotypic synergism in *Met* and *broad* double mutants, demonstrating JH-sensitive MET and BROAD interaction (BROAD protein accumulation is comparable to that of wild type flies, suggesting physical interaction with, rather than transcriptional regulation by *Met*), and providing a link between JH and 20E signaling (Wilson *et al.*, 2006).

In a hemimetabolous insect, *Oncopeltus fasciatus*, continuous *Broad* expression directs progressive development through nymphal instars (Erezyilmaz *et al.*, 2006). In Holometabola, *Broad* expression is confined to the prepupal stage, acting as a pupal specifier (Zhou & Riddiford, 2002). Loss of *Broad* expression, characteristic of the *npr1* mutant (*non-pupariating*; a deletion of the entire complementation group), results in the namesake phenotype of failure to enter the pupal program. Consequently, a restriction of *Broad* expression during this developmental stage may have contributed to the evolution of complete metamorphosis. During larval development in *D. melanogaster*, JH represses *broad*. At pupariation, exogenous JH induces a second wave of *broad* expression in the abdominal epidermis, resulting in the deposition of a second pupal cuticle (Zhou & Riddiford, 2002), demonstrating that the networks underlying these signaling mechanisms are complex.

In *T. castaneum*, methoprene exposure induces *Broad* expression and this upregulation is ablated upon *TcMet* knockdown. Therefore, *TcMet* is upstream of *Broad* in JH signaling in these beetles (Konopova & Jindra, 2008). *Krüppel homolog 1* (*Kr-h1*) is upstream of *Broad* in *D. melanogaster* JH signaling, where its expression in abdominal epidermis produces sternal bristle disruption similar to that seen following low dose JHA exposure (Minakuchi *et al.*,

2008). Genetic suppression of *TcKr-h1* induces precocious metamorphosis, similar to *TcMet* deficiency; *TcMet* knockdown in combination with JHA treatment demonstrated that *TcKr-h1* exists downstream of *TcMet* and upstream of *TcBroad* (Minakuchi *et al.*, 2009). Similarly, *Kr-h1* upregulation in newly eclosed *A. aegypti* females depends on *AaMet* expression (Zhu *et al.*, 2010). Therefore, the relationships among these genes are generally conserved within holometabolism evolution.

While *Kr-h1* also has demonstrated roles in JH-influenced social behavior of honeybees, it has not been reported whether *AmKr-h1* is under the transcriptional control of an *A. mellifera* *Met*-like protein. However, there is evidence for conservation in the sets of genes regulated by JH between flies and bees. This is perhaps unsurprising given the deep evolutionary conservation of these genetic mechanisms; *Kr-h1* and *Broad* expression profiles in two species of hemimetabolous thrips, whose life histories involve pupa-like, quiescent or non-feeding stages, are compatible with the expression profiles of *Broad* and *Kr-h1* in holometabolous insects (Minakuchi *et al.*, 2011).

Microarray data from *D. melanogaster* and *A. mellifera* identified a subset of conserved, JH-inducible genes (Li *et al.*, 2007). In the promoter region of 16 of the *D. melanogaster* orthologs, a conserved JH response element (JHRE) was identified. RNAi-driven reduction of the expression of two proteins identified as JHRE binders, *FKBP39* and *Chd64*, inhibits JHIII-induced expression of a reporter construct, suggesting their involvement in JH-dependent transcriptional machinery. Bitra and Palli (2009) demonstrated physical interaction of MET with both ECR and USP. Furthermore, column pulldown assays showed FKBP39 and CHD64 as binding partners of *D. melanogaster* ECR, USP, and MET, providing a more robust framework for a protein complex involving constituents of both JH and 20E signaling pathways (Li *et al.*, 2007). *FKBP39*, which is present at the onset of metamorphosis (Riddiford, 2008), is an inhibitor of autophagy in *D. melanogaster*; *FKBP39* overexpression precludes the developmental autolysis of larval fat body cells in wandering third instar larvae (Juhász *et al.*, 2007), a physiology shown to be partially dependent on MET/GCE regulation of caspase gene expression (Liu *et al.*, 2009). A role for GCE in any of these protein complexes has yet to be reported. *Chd64* is expressed during larval molts, but not in the third instar or during metamorphosis (Riddiford, 2008). Accordingly, putative regulatory complexes consisting of different combinations of these elements may assemble in a stage- or tissue-specific manner. Assembly of differential protein complexes in response to JH, 20E, or both could be a strategy for the tight regulation of the activities of these counteracting hormones.

The *Met*-like genes of *Tribolium* and *Drosophila* appear to act in similar genetic environments to regulate the expression of members of the 20E induced transcriptional cascade, including EcR (Riddiford *et al.*, 2010) and USP (Xu *et al.*, 2010), the heterodimeric components of the ecdysone receptor, various orphan nuclear receptors involved in 20E activity, and 20E-induced caspase genes involved in PCD (Liu *et al.*, 2009). Knockdown of seven nuclear receptors (E75, HR3, EcR, USP, SVP, FTZ-F1, and HR4) results in a significant reduction of vitellogenin production in *T. castaneum* (Xu *et al.*, 2010), a phenotype similar to that obtained via *TcMet* knockdown (Parthasarathy & Palli, 2009). The data presented in this section therefore strongly support for the action of *Met*-like genes as crucial to 20E/JH crosstalk.

## 5.2 Discovery of an evolutionarily conserved *Met* binding partner

Recent biochemical data from *A. aegypti* indicate that *AaMet* binds another bHLH PAS gene, *AaFISC*, and that this interaction requires a high JH titer. FISC is a coactivator of EcR/USP



(Chen, J.D., 2000; Zhu, *et al.*, 2006), providing yet another link between 20E and JH signaling. The authors also report that coexpression of *DmMet* or *Dmgce* with *DmTaiman* (the *D. melanogaster* *AaFISC* ortholog) in the presence of JH III induced reporter gene expression in L57 cells (Li *et al.*, 2011). Furthermore, this interaction has also been demonstrated in *T. castaneum* between *TcMet* and the FISC/TAI homolog, *TcSRC* (Steroid receptor coactivator; Zhang *et al.*, 2011). This gene has previously been implicated in metamorphic activity. *T. castaneum* larvae treated with SRC RNAi fail to achieve critical weight and consequently die before the larval-pupal transition (Bitra *et al.*, 2009). Therefore, MET interaction with FISC/SRC/TAIMAN underpins key transcriptional events of JH signaling throughout holometabolous insects.

Structure-function analyses performed using site-directed mutagenesis identified regions of MET that are necessary for homodimerization and GCE binding. Point mutations in the bHLH and PAS A domains (*Met<sup>1</sup>* and *Met<sup>tw3</sup>* alleles, respectively) had no effect on partner binding, whereas N- and C-terminal truncations, deletions in the HLH or PAS A domains, and a point mutation in the PAS B domain (*Met<sup>128</sup>* allele) all inhibited dimerization (Godlewski, *et al.*, 2006). Structure-function data for AaMET and AaFISC binding illustrates that the criticality of PAS domains for protein-protein interaction. Interestingly, two-hybrid assays showed that MET/FISC interaction increased when AaMET lacked a bHLH domain (Li *et al.*, 2011). Therefore, this domain is unnecessary for MET-FISC interaction, suggesting that the sole function of the *AaMet* bHLH may be in DNA binding. In contrast, deletion of the bHLH domain in FISC hindered the JH-induced interaction with AaMET.

Mixespression of *DmTaiman* in a variety of *Met/gce* genetic backgrounds will be valuable from both physiological and evolutionary perspectives. How do these proteins interact in the context of hormonal control of *D. melanogaster* development? Presumably, during larval development JH secreted from the CA inhibits MET and GCE interaction while promoting MET and GCE binding with TAIMAN. Are MET:TAI and GCE:TAI dimers functionally congruent in *D. melanogaster* or do these complexes preferentially regulate disparate target genes? Is the *Met*-like gene in *A. aegypti* and *T. castaneum* functionally analogous to *Met/gce* or are other, unidentified proteins involved? How has the interaction of these proteins changed during dipteran evolution following the origin of *Met*?

## 6. Conclusions

RT-PCR analysis with degenerate primers identified a single *Met*-like homolog in the genome of each of the three mosquito species, *Aedes aegypti*, *Anopheles gambiae*, and *Culex pipiens*. Likewise, a single *Met*-like ortholog exists in the beetle, *T. castaneum*, (Konopva & Jindra, 2007). Phylogenetic analysis and comparison of intron number and position in each of the identified mosquito genes indicates that the mosquito *Met* orthologs share higher sequence identity with *Dmgce* than *DmMet*, suggesting that *DmMet* arose from the duplication of an ancestral, *gce*-like gene in lower Diptera. To examine the evolutionary history of *Met* and *gce* within the Diptera, we mined the public *G. morsitans* EST library, recovering unique putative *Met* and *gce* orthologs in this fly, showing conservation of a *Met* homolog within the Schizophora. We also recently isolated a putative *gce* homolog from a Bombyliid, *Bombylius major* (A.A.B., unpublished). Taxonomically, this group of flies exists in the Asilomorpha, a paraphyletic sister taxon to the Muscomorpha within the dipteran infraorder, Brachycera. While this study is in its preliminary stages, only a single *Met*-like gene has thus far been obtained from this fly using degenerate PCR with cDNA and

genomic DNA templates, suggesting the possibility that the *Met/gce* duplication occurred within the Brachycera. *Met* function is evolutionarily conserved in Diptera; consistent with independent reports (Zhu *et al.*, 2010) we observed that RNAi-driven reduction of *AaMet* results in concomitant reduction of JH-inducible genes (Figure 4).

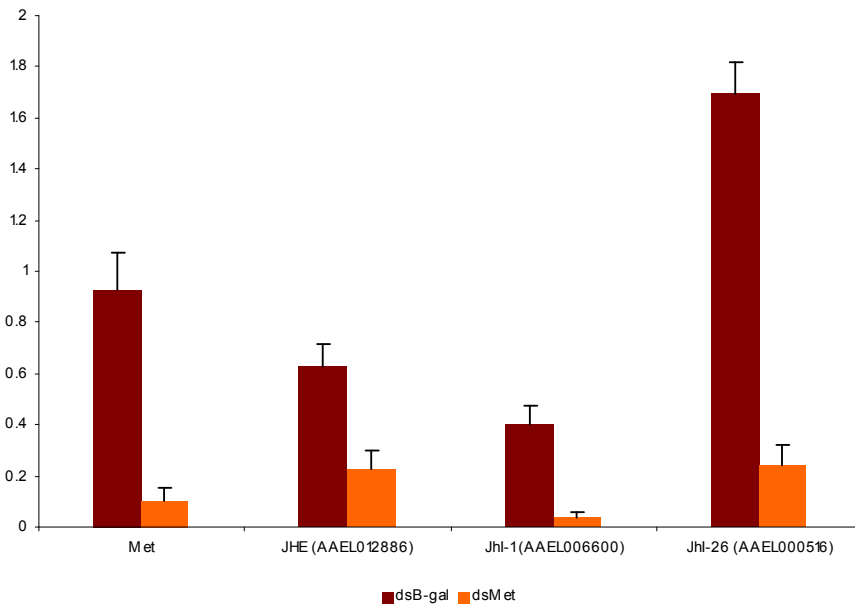


Fig. 4. Expression of three JH-inducible genes following RNAi-induced knockdown of *AaMet* (dsMet: orange bars) vs. controls (dsB-gal: red bars). *AaMet* reduction produced concomitant suppression of the *A. aegypti* homologs of *DmJHE*, *DmJhl-1*, and *DmJhl-26*.

Analysis of the nonsynonymous-to-synonymous substitution ratios (dN/dS) of *Met* and *gce* orthologs within the genus *Drosophila* indicates a substantial relaxation of selective constraint on the C-terminal half of *gce*, downstream of the functional domains. Conversely, nonsynonymous substitutions in the N-terminal half are stringently selected against. Depressed dN/dS values across the *Met* coding sequence indicate strong selective constraint over the entire open reading frame (Baumann *et al.*, 2010b).

RT-PCR analysis of selected *D. melanogaster* tissues shows that *gce* is generally co-expressed with *Met* in known JH target tissues, including ovary and MAG. Overexpression of *gce* in a *Met* mutant background results in a dramatic enhancement of methoprene-conditional toxic and morphogenetic defects, similar to those seen in wild type (*Met*<sup>+</sup>) flies after methoprene exposure. *Met* mutant flies overexpressing *gce* show rescue of a non-conditional adult phenotype, that of defective development of posterior facets in the compound eye. Our results therefore support the notion of functional redundancy that has been hypothesized to account for *Met*<sup>27</sup> viability flies. On the other hand, we have also shown that these paralogs have undergone evolutionary subfunctionalization since their origin; *gce* overexpression

fails to rescue the phenotypes of deficient oogenesis or reduced male courtship characteristic of *Met* adults, showing that *Met* has co-opted the role as the mediator of JH-regulated reproductive functions in *Drosophila*.

RNAi-driven reduction of *gce* expression from either an *actin* or *tubulin* promoter demonstrates that unlike *Met*, *gce* is a vital gene. *Gce* underexpression in both *Met*<sup>+</sup> and *Met* genetic backgrounds results in lethality. *Met*<sup>+</sup> ; *UAS-gce-dsRNA* / *tubulin-GAL4* homo-/hemizygotes do not survive to adulthood, and die primarily at the pharate adult stage, while the same *gce* RNAi construct expressed in a *Met* mutant background shifts lethality to early pupae (Baumann *et al.*, 2010a).

### 6.1 Directions

Previously, USP was proposed as a candidate JH receptor (Jones, *et al.*, 2001). Yet, USP only binds JH with micromolar affinity, requiring a hormone concentration that exceeds endogenous titers by orders of magnitude (Bownes & Rembold, 1987). It is now known that USP binds methyl farnesoate (MF), a precursor in the biological synthesis of JH III (Figure 1), with nanomolar affinity both in *D. melanogaster* and in *A. aegypti* (Jones *et al.*, 2006; Jones *et al.*, 2010). Recent studies on natural farnesoid derivatives including MF, JH III, and JHB3 (the main farnesoid secretion product of dipteran ring glands cultured *in vitro*) have teased out the relative activities of each of these compounds during development in a series of biological assays. Two recent studies have demonstrated that the activity series of these three compounds changes during development. Dietary MF and JH III (MF > JH III) were both more active than JHB3 in delaying larval attainment of the wandering stage. In contrast, JH III applied to prepupae (white puparial assay; Riddiford & Ashburner, 1991) showed much higher activity than MF or JHB3 in blocking adult eclosion (Jones *et al.*, 2010; Harshman *et al.*, 2010).

Topical application or dietary provision of these compounds adds to endogenous hormone titers. Therefore, just as USP binds JH III at concentrations exceeding physiological levels, it is possible that MET nonspecifically binds MF or JHB3 under these conditions. It is an intriguing proposition that MET and USP, which interact both with each other and JHRE binding proteins (Bitra & Palli, 2009), may partner in a stage-specific manner throughout development in response to a fluctuating mélange of methyl farnesoids. Does GCE participate in the assembly of the molecular machinery that facilitates the crosstalk between JH and 20E signaling? This protein has been largely ignored in studies regarding the molecular interaction of these hormones. Further, it is unknown whether GCE binds any of the farnesoid products of the CA. There appears to be a correlation between the presence of paralogous *Met*-like genes and multiple JH isoforms in higher Diptera. If each of the farnesoids JH III, MF, and JHB3 indeed has a unique receptor protein, the possibility arises that GCE fills the role of JHB3 binder. Or perhaps in MET/GCE dimers, MET is the sole ligand binder, while GCE and MET are both necessary for target gene transcription. Clearly, further functional characterization of GCE is necessary to unravel the mechanisms through which JH signaling has evolved from the basal holometabola to the most evolutionarily diverged insects, the higher Diptera.

## 7. Acknowledgments

We extend thanks to Drs. John Freudenstein and H. Lisle Gibbs for their guidance in our various evolutionary analyses. Additionally, we would like to recognize Dr. Shaoli Wang

for her expertise and contributions to much of the work presented herein, and for her mentorship of AAB. Work presented in this text was supported by NIH grant AI052290 to T.G.W.

## 8. References

- Ashburner, M., 2007. *Drosophila* Genomes by the Baker's Dozen. Preface. *Genetics*. 177:1263-1268.
- Ashburner, M., Chihara, C., Meltzer, P., Richards, G. 1974. Temporal control of puffing activity in polytene chromosomes. *Cold Spring Harbor Symposia on Quantitative Biology*. 38:655-662.
- Ashok M., Turner, C., Wilson, T.G., 1998. Insect juvenile hormone resistance gene homology with the bHLH-PAS family of transcriptional regulators. *Proceedings of the National Academy of Sciences, USA*. 95:2761-2766.
- Barry, J., Wang, S., Wilson, T.G., 2008. Overexpression of *Methoprene-tolerant*, a *Drosophila melanogaster* gene that is critical for juvenile hormone action and insecticide resistance. *Insect Biochemistry and Molecular Biology*. 3:346-353.
- Baumann, A., Fujiwara, Y., Wilson, T.G. 2010a. Evolutionary divergence of the paralogs *Methoprene tolerant* (*Met*) and *germ cell expressed* (*gce*) within the genus *Drosophila*. *Journal of Insect Physiology*. 56:1445-1455.
- Baumann, A., Wilson, T.G., Barry, J., Wang, S. 2010b. Juvenile hormone action requires paralogous genes in *Drosophila melanogaster*. *Genetics*. 185:1327-1336.
- Bertone, M.A., Wiegmann, B.M. 2009. *Diptera* (true flies). In: *Hedges, S.B., Kumar, S. (Eds.), The Timetree of Life*. Oxford, New York. 270-277.
- Bitra, K., Palli, S.R. 2009. Interaction of proteins involved in ecdysone and juvenile hormone signal transduction. *Archives of Insect Biochemistry and Physiology*. 70:90-105.
- Bitra, K., Tan, A., Downling, A., Palli, S.R. 2009. Functional characterization of PAS and HES family bHLH transcription factors during the metamorphosis of the red flour beetle, *Tribolium castaneum*. *Gene*. 448:74-87.
- Bouchard B.L., Wilson, T.G. 1987. Effects of sublethal doses of Methoprene on reproduction and longevity of *Drosophila melanogaster* (Diptera: Drosophilidae). *Journal of Economic Entomology*. 80:317-21.
- Bownes, M., Rembold, H. 1987. The titure of juvenile hormone during the pupal and adult stages of the life cycle of *Drosophila melnoagaster*. *European Journal of Biochemistry*. 164:709-712.
- Brand, A.H., Perrimon, N. 1993. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development*. 118:401-415.
- Bylemans, D., Borovsky, D., Ujvary, I., DeLoof, A. 1998. Biosynthesis and regulation of juvenile hormone III, juvenile hormone III bisepoxide, and methyl farnesoate during the reproductive cycle of the grey flesh fly, *Neobellieria* (Sarcophaga) *bullata*. *Archives of Insect Biochemistry and Physiology* 37:248-256.
- Chen, J.D. 2000. Steroid/nuclear receptor coactivators. *Vitamins and Hormones*. 58:391-448.
- Cornel, A., Stanich, M., McAbee, R., Mulligan, F., 2002. High level methoprene resistance in the mosquito *Ochlerotatus nigromaculis* Ludlow in Central California. *Pest Management Science*. 58:791-798.
- Cornel, A.J., Stanich, M.A., Farley, D., Mulligan III, F.S., Hyde, G. 2000. Methoprene tolerance in *Aedes nigromaculis* in Fresno County, California, *Journal of American Mosquito Control Association*. 16:223-228.

- Crews, S.T., 2003. PAS Proteins, Regulators and Sensors of Development and Physiology. Boston, Kluwer.
- Dame, D.A., Wichterman, G.J., Hornby, J.A., 1998. Mosquito *Aedes taeniorhynchus* resistance to methoprene in an isolated habitat. *Journal of the American Mosquito Control Association*. 14:200-203.
- Davis, R.E., Kelly, T.J., Masler, E.P., Fescemyer, H.W., Thyagaraja, B.S., Borkovec, A.B. 1990. Hormonal control of vitellogenesis in the gypsy moth, *Lymantria dispar* (L.): suppression of haemolymph vitellogenin by the juvenile hormone analogue, methoprene. *Journal of Insect Physiology*, 36:231-238.
- Denlinger, D.L., 1985. Hormonal control of diapause. In: Kerkut, G.A., Gilbert, L.I. (Eds.), *Comprehensive Insect Physiology, Biochemistry and Pharmacology*, vol. 8. Pergamon, New York. 37-84.
- Dubrovsky E.B., Dubrovskaya V.A., Bilderback, A. L., Berger, E.M. 2000. The isolation of two juvenile hormone-inducible genes in *Drosophila melanogaster*. *Developmental Biology*. 224:486-495.
- Dubrovsky, E.B., Dubrovskaya, V.A., Berger, E.M., 2002. Juvenile hormone signaling during oogenesis in *Drosophila*. *Insect Biochemistry and Molecular Biology*. 32:1555-1565.
- Dubrovsky, E.B., Dubrovskaya V.A., Berger E.M. 2004. Hormonal regulation and functional role of *Drosophila* E75A orphan nuclear receptor in the juvenile hormone signaling pathway. *Developmental Biology*. 268:258-270.
- Erezylmaz, D.F., Riddiford, L.M., Truman, J.W. 2006. The pupal specifier *broad* directs progressive morphogenesis in a direct-developing insect. *Proceedings of the National Academy of Sciences, USA*. 103:6925-6930.
- Flatt, T., Heyland, A., Rus, F., Porpiglia, E., Sherlock, C., Yamamoto, R., Garbuzov, A., Palli, S.R., Tatar, M., Silverman, N. 2008. Hormonal regulation of the humoral innate immune response in *Drosophila melanogaster*. *Journal of Experimental Biology*. 211:2712-24.
- Flaveny, C.A., Murray, I.A., Perdew, G.H., 2010. Differential gene regulation by the human and mouse *aryl hydrocarbon receptor*. *Toxicological Sciences*. 114:217-225.
- Godlewski, J., Wang, S., and Wilson, T.G., 2006. Interaction of bHLH-PAS proteins involved in juvenile hormone reception in *Drosophila*. *Biochemical and Biophysical Research Communications*. 3424:1305-1311.
- Harshman, L.G., Song, K.D., Casas, J., Schuurmans, A., Kuwano, E., Kachman, S.D., Riddiford, L.M., Hammock, B.D. 2010. Bioassays of compounds with potential juvenoid activity on *Drosophila melanogaster*: juvenile hormone III, bisepoxide juvenile hormone III and methyl farnesoates. *Journal of Insect Physiology*. 56:1465-1470.
- Huang, Z. J., Edery, I., Rosbash, M. 1993. PAS is a dimerization domain common to *Drosophila Period* and several transcription factors. *Nature*. 364: 259-262.
- Juhász, G., Puskás, L.G., Komonyi, O., Erdi, B., Maróy, P., Neufeld, T.P., Sass, M. 2007. Gene expression profiling identifies FKBP39 as an inhibitor of autophagy in larval *Drosophila* fat body. *Cell Death and Differentiation*. 14:1181-1190.
- Kethidi, D.R., Xi, Z., Palli, S.R. 2005. Developmental and hormonal regulation of juvenile hormone esterase gene in *Drosophila melanogaster*. *Journal of Insect Physiology*. 51:393-400.
- Konopova, B., Jindra, M., 2007. Juvenile hormone resistance gene *Methoprene-tolerant* controls entry into metamorphosis in the beetle *Tribolium castaneum*. *Proceedings of the National Academy of Sciences, USA*. 10419:10488-10493.

- Kosakovsky Pond, S., Frost, S., 2005. DataMonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*. 21:2531-2533.
- Kosakovsky Pond, S., Muse, S., 2005. HyPhy, Hypothesis Testing Using Phylogenies. In: Nielsen, R (Ed.), *Statistical Methods in Molecular Evolution*. Springer, New York, pp. 125-182.
- Li, Y., Zhang, Z., Robinson, G.E., Palli, S.R., 2007. Identification and characterization of a juvenile hormone response element and its binding proteins. *Journal of Biological Chemistry*. 52:37605-37617.
- Li, Z., Cheng, D., Wei, L., Zhao, P., Shu, X., Tang, L., Xiang, Z., Xia, Q. 2010. The silkworm homolog of *Methoprene-tolerant* (*Met*) gene reveals sequence conservation but function divergence. *Insect Science*, 17:313-324.
- Liu, Y., Sheng, Z., Liu, H., Wen, D., He, Q., Wang, S., Shao, W., Jiang, R.J., An, S., Sun, Y., Bendena, W.G., Wang, J., Gilbert, L.I., Wilson, T.G., Song, Q., Li, S., 2009. Juvenile hormone counteracts the bHLH-PAS transcription factors MET and GCE to prevent caspase-dependent programmed cell death in *Drosophila*. *Development*. 13612:2015-2025.
- Jones, G., Jones, D., Li, X., Tang, L., Ye, L., Teal, P., Riddiford, L., Sandifer, C., Borovsky, D., Martin, J. 2010. Activities of natureal methyl farnesoids on pupariation and metamorphosis of *Drosophila melanogaster*. *Journal of Insect Physiology*. 56:1456-1464.
- Jones, G., Jones, D., Teal, P., Sapa, A., Wozniak, M. 2006. The *retinoid-X receptor* ortholog, *ultraspiracle*, binds with nanomolar affinity to an endogenous morphogenetic ligand. *FEBS Journal*. 273:4983-4996
- Jones, G., Wozniak, M., Chu, Y.X., Dhar, S., Jones, D. 2001. Juvenile hormone III-dependent conformational changes of the nuclear receptor *ultraspiracle*. *Insect Biochemistry and Molecular Biology*. 32:33-49.
- Konopova, B., Jindra, M. 2008. *Broad-Complex* acts downstream of *Met* in hjuvenile hormone signalling to coordinate primitive holometabolan metamorphosis. *Development*. 135:559-568.
- MacCarthy, T., Bergman, A., 2007. The limits of subfunctionalization. *BMC Evolutionary Biology*. 7:213-226.
- Madhavan, K., 1973. Morphogenetic effects of juvenile hormone and juvenile hormone mimics on adult development of *Drosophila*. *Journal of Insect Physiology*. 19:441-453.
- Minakuchi, C., Namiki, T., Shinoda, T. 2009. *Krüppel homolog 1*, an early juvenile hormone-response gene downstream of *Methoprene-tolerant*, mediates its anti-metamorphic action in the red flour beetle, *Tribolium castaneum*. *Developmental Biology*. 325:341-350.
- Minakuchi, C., Tanaka, M., Miura, K., Tanaka, T. 2011. Developmental progile and hormonal regulation of the transcription factors *broad* and *Krüppel homolog 1* in hemimetabolous thrips. *Insect Biochemistry and Molecular Biology*. 41:125-134.
- Minakuchi, C., Zhou, X., Riddiford, L.M. 2008. *Krüppel homolog 1* (*Kr-h1*) mediates juvenile hormone action during metamorphosis of *Drosophila melanogaster*. *Mechanisms of Development*. 125:91-105.
- Minkhoff III, C., Wilson, TG. 1992. The competitive ability and fitness components of the *Methoprene-tolerant* (*Met*) *Drosophila* mutant resistant to juvenile hormone analog insecticides. *Genetics*, 131:91-97.
- Miura, K., Oda, M., Makita, S., Chinzei, Y., 2005. Characterization of the *Drosophila Methoprene-tolerant* gene product: juvenile hormone binding and ligand-dependent gene regulation. *FEBS Journal*. 2725:1169-1178.

- Moore, A.W., Barbel, S., Jan, L.Y., Jan, Y.N., 2000. A genomewide survey of basic helix-loop-helix factors in *Drosophila*. Proceedings of the National Academy of Sciences, USA. 9719:10436-10441.
- Parthasarathy, R. and Palli, S. R. 2009. Molecular analysis of juvenile hormone action in controlling the metamorphosis of the red flour beetle, *Tribolium castaneum*. Archives of Insect Biochemistry and Physiology. 70:57-70.
- Parthasarathy, R., Tan, A., Bai, H., Palli, S.R., 2008. bHLH-PAS family transcription factor *Methoprene-tolerant* plays a key role in JH action in preventing the premature development of adult structures during larval-pupal metamorphosis. Mechanisms of Development. 1257:601-616.
- Parthasarathy, R., Sun, Z., Bai, H., Palli, S.R. 2010. Juvenile hormone regulation of vitellogenin synthesis in the red flour beetle, *Tribolium castaneum*. Insect Biochemistry and Molecular Biology. 40:405-414.
- Postlethwait, J.H. 1974. Juvenile hormone and the adult development of *Drosophila*. Biological Bulletin. 147:119-135.
- Postlethwaite and Weiser, 1973. Vitellogenesis induced by juvenile hormone in the female sterile mutant apterous-four in *Drosophila melanogaster*, Nature New Biology. : 284-285.
- Pursley, S., M. Ashok, Wilson, T.G. 2000. Intracellular localization and tissue specificity of the *Methoprene-tolerant* (*Met*) gene product in *Drosophila melanogaster*. Insect Biochemistry and Molecular Biology. 30: 839-845.
- Ramadoss, P., Perdew, G.H., 2005. The transactivation domain of the *Ah* receptor is a key determinant of cellular localization and ligand-independent nucleocytoplasmic shuttling properties. Biochemistry. 4433:11148-11159.
- Richard, D.S., Applebaum, S.W., Sliter, T.J., Baker, F.C., Schooley, D.A., Reuter, C.C., Henrich, V.C., Gilbert, L.I. 1989. Juvenile hormone bisepoxide biosynthesis in vitro by the ring gland of *Drosophila melanogaster*: a putative juvenile hormone in the higher Diptera. Proceedings of the National Academy of Sciences, USA. 86:1421-1425.
- Riddiford, L., 2008. Juvenile hormone action, a 2007 perspective. Journal of Insect Physiology. 54:895-901.
- Riddiford, L.M. and Ashburner, M. 1991. Effects of juvenile hormone mimics on larval development and metamorphosis of *Drosophila melanogaster*. General and Comparative Endocrinology. 82:172 -183.
- Riddiford, L.M., 1994. Cellular and molecular actions of juvenile hormone. I. General considerations and premetamorphic actions. Advances in Insect Physiology. 24:213-274.
- Riddiford, L.M., Truman, J.W., Mirth, C.K., Shen, Y.C. 2010. A role for juvenile hormone in the prepupal development of *Drosophila melanogaster*. Development. 137:1117-1126.
- Shemshedini, L., Wilson, T.G. 1990. Resistance to juvenile hormone and an insect growth regulator in *Drosophila* is associated with an altered cytosolic juvenile hormone binding protein. Proceedings of the National Academy of Sciences, USA. 87: 2072-2076.
- Soller, M., Bownes, M., Kubli, E. 1999. Control of oocyte maturation in sexually mature *Drosophila* females. Developmental Biology. 208:337-51.
- Srivastava, U.S., Srivastava, R.C. 1983. Juvenoid-induced supernumerary larval instars in certain stored grain insects. Proceedings: Animal Sciences. 92:263-276.
- Tamura, K., Subramanian, S., and Kumar, S., 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Molecular Biology and Evolution. 21:36-44.
- Tomkins, L. 1990. Effects of the *Apterous* mutation on *Drosophila melanogaster* males' courtship. Journal of Neurogenetics. 6:221-227.

- Truman, J.W., Riddiford, L.M. 2002. Endocrine insights into the evolution of metamorphosis in insects. *Annual Review of Entomology*. 47:467-500.
- Wang S., Baumann, A., Wilson, T.G., 2007. *Drosophila melanogaster* Methoprene- tolerant (*Met*) gene homologs from three mosquito species, members of PAS transcriptional factor family. *Journal of Insect Physiology*. 533:246-253.
- Williams, C. M. 1956. The juvenile hormone of insects. *Nature*. 178: 212-213.
- Williams, C.M., 1967. Third-generation pesticides. *Scientific American*. 217:13-17.
- Wilson TG, Wang S, Beno M, Farkas R. 2006. Wide mutational spectrum of a gene involved in hormone action and insecticide resistance in *Drosophila melanogaster*. *Molecular Genetics and Genomics*. 2763:294-303.
- Wilson, T.G. 1982. A correlation between juvenile hormone deficiency and vitellogenic oocyte degeneration in *Drosophila melanogaster*. *Wilhelm Roux's Archives Entwicklungsmechanik der Organismen*. 191:257-263.
- Wilson, T.G., and Ashok, M., 1998. Insecticide resistance resulting from an absence of target-site gene product. *Proceedings of the National Academy of Sciences, USA*. 9524:14040-14044.
- Wilson, T.G., and Fabian, J. 1986. A *Drosophila melanogaster* mutant resistant to a chemical analog of juvenile hormone. *Developmental Biology*. 118:190-201.
- Wilson, T.G., DeMoor, S., Lei, J. 2003. Juvenile hormone involvement in *Drosophila melanogaster* male reproduction as suggested by the *Methoprene-tolerant*<sup>27</sup> mutant phenotype. *Insect Biochemistry and Molecular Biology*. 3312:1167-1175.
- Wilson, T.G., Yerushalmi, Y., Donnell, D.M., Restifo, L.L. 2006. Interaction between hormonal signaling pathways in *Drosophila melanogaster* as revealed by genetic interaction between *Methoprene-tolerant* and *Broad-Complex*. *Genetics* 172:253-264.
- Xu, J., Tan, A., Palli, S.R. 2010. The function of nuclear receptors in regulation of female reproduction and embryogenesis in the red flour beetle, *Tribolium castaneum*. *Journal of Insect Physiology*. 56:1471-1480.
- Yamamoto, K., Chadarevian, A., Pellegrini, M. 1988. Juvenile hormone action mediated in male accessory glands of *Drosophila* by calcium and kinase C. *Science*. 239:916-919.
- Yang, X.H., Liu, P.C., Zheng, W.W., Zhao, X.F. 2011. The juvenile hormone analogue methoprene up-regulates the Ha-RNA-Binding protein. *Molecular and Cellular Endocrinology*. 333:172-180.
- Zhang, Z., Xu, J., Sheng, Z., Sui, Y., Palli, S.R. 2011. *Steroid Receptor Co-activator* is required for juvenile hormone signal transduction through a bHLH-PAS transcription factor, *Methoprene tolerant*. *Journal of Biological Chemistry*. 286: 8437-8447.
- Zhou, B., Hiruma, K., Shinoda, T., Riddiford, L.M. 1998. Juvenile hormone prevents ecdysteroid-induced expression of broad complex RNAs in the epidermis of the tobacco hornworm, *Manduca sexta*. *Developmental Biology*. 203:233-44.
- Zhou, X., Riddiford, L. M. 2002. *Broad* specifies pupal development and mediates the 'status quo' action of juvenile hormone on the pupal-adult transformation in *Drosophila* and *Manduca*, *Development*. 129:2259-2269.
- Zhu, J., Busche, J.M., Zhang, X. 2010. Identification of juvenile hormone target genes in the adult female mosquitoes. *Insect Biochemistry and Molecular Biology*. 40:23-29.
- Zhu, J., Chen, L., Sun, G., Raikhel, A.S. 2006. The competence factor beta Ftz-F1 potentiates ecdysone receptor activity via recruiting a p160/SRC coactivator *Molecular and Cellular Biology*. 26:9402-9412.



# Gene Duplication and Subsequent Differentiation of Esterases in Cactophilic *Drosophila* Species

Rogério P. Mateus<sup>1</sup>, Luciana P. B. Machado<sup>1</sup> and Carlos R. Ceron<sup>2</sup>

<sup>1</sup>Universidade Estadual do Centro-Oeste – UNICENTRO

<sup>2</sup>Instituto de Biociências, Letras e Ciências Exatas – IBILCE,  
Universidade Estadual Paulista –UNESP  
Brazil

## 1. Introduction

Phytophagous insects are excellent model systems to study the genetic and ecological bases of adaptation and population differentiation because the host plant constitutes an immediate environmental factor that can affect the early stages of the life cycle (Matzkin, 2005; Matzkin et al., 2006). New host plant exploitation can result in genetic and biochemical adjustments to the new resource and to chemically distinct niches, which can include potentially toxic compounds, new mating environments, parasitoids, bacteria and fungi (Kircher, 1982; Fogleman & Abril, 1990; Via, 1990; Fogleman & Danielson, 2001). These adjustments are the result of a number of physiological changes, including those related to biochemical systems associated with adaptation to the new environment.

The species of the *Drosophila repleta* group occupy different habitats, but their common feature is that they are phytophagous; that is, they lay eggs in rotting cacti cladodes. The developing larvae feed on the yeast that are part of the rotting process (Starmer & Gilbert, 1982; Pereira et al., 1983; Starmer et al., 1986), according to the cactus-*Drosophila*-yeast system; therefore, they are considered specialists. However, adults are generalists because they visit other food sources in their environment (Morais et al., 1994). This ecological specificity of cactophilic *Drosophila* directly influences species distribution, as they are always associated with the host cactus distribution (Tidon-Sklorz & Sene, 1995; Manfrin & Sene, 2006; Mateus and Sene, 2007).

*Drosophila* has been used as a research model for more than a century, and the first report of gene duplication was described by Bridges for the *Bar* locus in *D. melanogaster* over 70 years ago (Bridges, 1936). Since that time, mainly after the advent of biochemical and molecular biology techniques, several other examples of duplicated genes have been presented, and pathways of evolution by gene duplication have been proposed (for example, Stephens, 1951; Nei, 1969). These pathways were thoroughly discussed in 1970 in Ohno's book "Evolution by gene duplication" (Ohno, 1970). Subsequently, several other works have reviewed the mechanisms and roles of gene duplication in the evolutionary process (A. Wagner, 2002; Kondrashov et al., 2002; and Zhang, 2003).

Currently, the genomes of twelve *Drosophila* species have been completely sequenced (Tweedie et al., 2009), but many aspects of the functional divergence of the products of a

gene duplication event cannot be answered through this method alone. A deeper investigation of genetic differentiation after duplication is possible through molecular and biochemical approaches. These approaches are extremely important because gene duplication followed by functional divergence has been considered the primary mechanism of molecular evolution (Lewis, 1951; Ohno, 1970). Analyses of isozymes have been crucial in this process because they provide, along with cytological studies, evidence for the frequent occurrence of gene duplication during the evolutionary process (Gottlieb, 1982; Hart, 1983). Esterase is a polymorphic group of isozymes that play important biochemical roles in insects. This group is composed of a heterogeneous set of hydrolytic enzymes that are widely distributed among organisms and that catalyze the hydrolysis of esters, peptides, amides and halides (Walker & Mackness, 1983). They are involved in digestive (Argentine & James, 1995) and reproductive processes (Karotam et al., 1993), the degradation of insecticides (Feyereisen, 1995) and female sex pheromones after male recognition (Vogt & Riddiford, 1981) and in the regulation of juvenile hormone levels (Gu & Zera, 1994). In *Drosophila*, esterases make up a diverse set of enzymes (G. B. Johnson, 1973, 1974), and gene duplication has been used as one explanation for their evolution (Zouros et al., 1982; Pen et al., 1990; Mateus et al., 2009).

Several studies on the changes in esterase activity during development in species of the *D. repleta* group have detected two main  $\beta$ -esterases that show different tissue-specific and temporal expression patterns (Zouros et al., 1982; East, 1982; Pen et al., 1984; Pen et al., 1986a, 1986b; Pen et al., 1990; Mateus et al., 2009). One esterase, named EST-4, is present only in later third instar larvae and early pupae and has a high concentration in the carcass. The other esterase, named EST-5, is present throughout the insect life cycle and occurs predominantly in hemolymph and the fat body (Zouros et al., 1982).

According to Zouros et al. (1982), who studied these enzymes in *D. mojavensis* and *D. arizonae*, the most likely hypothesis is that these enzymes are products of a gene duplication as old as the *D. repleta* group that diverged later regarding their patterns of tissue-specific and temporal expression. This hypothesis was suggested because these enzymes show interlocus heterodimers, different patterns of expression (tissue and temporal) and 82% identity in the N-terminal amino acid sequence (Pen et al., 1986a; Pen et al., 1990). More recently, Robin et al. (2009) demonstrated that these enzymes are encoded by two genes that are products of a gene duplication, *Est-2a* (EST-5) and *Est-2c* (EST-4), in *D. mojavensis*.

In this study, we investigated several genetic and biochemical features of EST-4 and EST-5 in six species of the *D. repleta* group, three belonging to the *D. mulleri* cluster (*Drosophila mulleri*, *D. aldrichi* and *D. wheeleri*) and three belonging to the *D. mojavensis* cluster (*D. mojavensis*, *D. arizonae* and *D. navojoa*) of the *D. mulleri* complex, as well as hybrids from crosses involving some of these species. We aimed to establish the biochemical and genetic differences among and possible physiological roles for these enzymes in the metabolic processes of this group of *Drosophila* species.

## 2. Materials and methods

### 2.1 Species

The materials used in this study included laboratory stocks of six *Drosophila* species (*D. arizonae* - AR, *D. mojavensis* - MO, *D. navojoa* - NA, *D. mulleri* - MU, *D. aldrichi* - AL and, *D. wheeleri* - WH) and two homozygote line stocks, one for the EST-5 "slow" allele of *D. mulleri* (MU-S) and another for the EST-4 "fast" and EST-5 "slow" alleles of *D. navojoa* (NA-FS). All

stocks were multifemales, except for the NA-FS stock, which was isofemale. The laboratory stocks were obtained from Prof. Dr. Hermione E. M. C. Bicudo (Department of Biology, IBILCE/UNESP, São José do Rio Preto, Brazil), who brought them to Brazil from stocks of the Genetics Foundation (University of Texas, Austin, TX, US). The two line stocks were prepared from laboratory stocks through endogamic crosses.

All laboratory and line stocks were maintained as mass cultures at a constant temperature of 20°C ±1°C in culture vials with standard banana medium. The origin of each stock is listed in Table 1.

Species	Codes	Locality
<i>Drosophila mulleri</i>	MU	Guayalejo, Tamazunchale, México
<i>Drosophila aldrichi</i>	AL	Austin, Texas, US
<i>Drosophila wheeleri</i>	WH	Arroyo Solloro, Baja California, México
<i>Drosophila mojavensis</i>	MO	Baja California, México
<i>Drosophila arizonae</i>	AR	Guayalejo, Tamazunchale, México
<i>Drosophila navojoa</i>	NA	Navojoa, México

Table 1. List of analyzed species, with the respective codes and original localities of the stocks (all stocks were obtained from the Genetics Foundation, University of Texas, Austin, TX, US).

## 2.2 Obtaining late third instar larvae and adult flies for electrophoresis

Late third instar larvae and adult flies were collected directly from the vials and immediately frozen at -20° C for further electrophoretic analyses. The larvae in that phase show yellowish spiraculum and maximum EST-4 activity. Female adult flies were collected at 5-10 days old and were used in electrophoresis for comparative analysis.

## 2.3 Obtaining hybrids

Mass crossings in both directions were performed in population boxes (16 cm<sup>3</sup>), using 200 couples, between NA-FS and MU-S, NA-FS and MO, NA-FS and AR and MU-S and MO. After setting up a cross, the courtship behavior was observed for 10 minutes, as described by Markow (1981). The culture media were placed in Petri plates at the bottom of the boxes and were changed every three days. After every plate change, the plates were inspected to detect eggs. The plates were maintained at a constant temperature of 20°C ±1°C until late third instar hybrid larvae were observed. These larvae were obtained directly from the plates and frozen at -20°C for further electrophoretic analyses.

## 2.4 Esterase detection

Esterase detection was performed using 10% polyacrylamide gel electrophoresis (PAGE), adapted by Ceron (1988) from Davis (1964) and Laemmli (1970). All samples were prepared in 25 µL of 0.1 M Tris-HCl (pH 8.8) buffer containing 10% glycerol, where 10 µL was used in the gels. After electrophoresis, all gels were soaked in 0.1 M phosphate buffer (pH 6.2) for 1 hour at 25°C. After this period, the gels were stained in solution containing 100 mL of phosphate buffer, 10 ml of n-propanol and 120 mg of Fast Blue RR Salt, where 40 mg of α-naphthyl acetate and 30 mg of β-naphthyl acetate, previously dissolved in 2 ml of acetone, were added. After approximately 1 hour, the staining reactions were stopped in a solution of acetic acid:ethanol:water (1:2.5:6.5 by v:v:v). Because the esterases hydrolyze substrates

differently, the bands in the gel stain differently: black when they hydrolyze  $\alpha$ -naphthyl acetate, red when they hydrolyze  $\beta$ -naphthyl acetate and magenta when they hydrolyze both  $\alpha$ - and  $\beta$ -naphthyl acetates. Polyacrylamide gels were air dried at room temperature using gelatin and cellophane wound slab gels in an embroidering hoop (Ceron et al., 1992).

## 2.5 Characterization of esterases using inhibitors

Malathion, phenylmethylsulfonyl fluoride (PMSF), eserine sulfate, copper sulfate ( $\text{CuSO}_4$ ), iodoacetamide (IAC), trans-epoxysuccinyl-L-leucyl-amido(4-guanidino) butane (E-64), p-chloromercuribenzoate (pCMB) and mercuric chloride ( $\text{HgCl}_2$ ) were used as specific inhibitors, all in 1 mM concentrations (with the exception of E-64, which was used at a concentration of 5 mM) in the soaking and staining solution.

## 2.6 Determination of isoelectric point (I.P.)

The I.P. was determined in 10% PAGE containing 5% ampholyte solution (Sigma). The first ampholyte formed a pH gradient between 3.0 and 10.0 after 1 hour of constant 100 V pre-focusing. This experiment was performed to verify the best gradient to determine the I.P. values of all enzymes in all species. After this verification, another ampholyte was used that formed a pH gradient between 6.0 and 8.0 after 1 hour of constant 100 V pre-focusing. In both cases, ampholyte solutions were added before gel polymerization. Samples of the six *Drosophila* species and of the I.P. marker (hemoglobin) were prepared in a 10% glycerol in water solution. Esterase isoelectric focusing was performed under constant 100 V conditions in the power supply for 3 hours. After focusing, the gels were soaked in buffer for 1 hour, followed by esterase staining for the same period, as described in section 2.5. Following esterase identification, the gels were stained for total protein with Coomassie Blue G250 overnight. The staining reaction was stopped, and the gels were dried as described in section 2.5. The I.P. was estimated by comparing the positions of EST-4 and EST-5 with the position of human hemoglobin (I.P. = 7.1) after focusing.

## 2.7 Molecular weight (MW) estimation

The MW estimation was performed using the method adapted by Mateus et al. (2009) from Hedrick and Smith (1968). The following standard MW proteins were used: myoglobin (17.8 kD), soybean trypsin inhibitor (24 kD), carbonic anhydrase (29 kD), ovalbumin (45 kD), human serum albumin (66 kD) and phosphorylase-b (97.4 kD). All graphics were constructed using Microcal Origin software, version 3.5 (Scientific and Technical Graphics in Windows – copyright 1991 – 1994, Microcal Software Inc.).

# 3. Results

## 3.1 Esterase pattern

Figure 1 shows the esterase patterns of larvae and adults (females) from six *Drosophila* species. For all species, EST-4 always migrated slower than EST-5. The *D. navojoa* stock was the only one that had more than one band for EST-4. EST-4 was more strongly stained than EST-5 in *D. mulleri*, *D. aldrichi* and *D. wheeleri* (Figure 2A and B). The opposite was observed for *D. arizonae*, with EST-5 more strongly stained than EST-4 (Figure 2A). Differences in the staining intensity among EST-4 bands were also observed, with the *D. mojavensis* cluster species (*D. mojavensis*, *D. arizonae* and *D. navojoa*) showing fainter bands than the species of the *D. mulleri* cluster (*D. mulleri*, *D. aldrichi* and *D. wheeleri*).

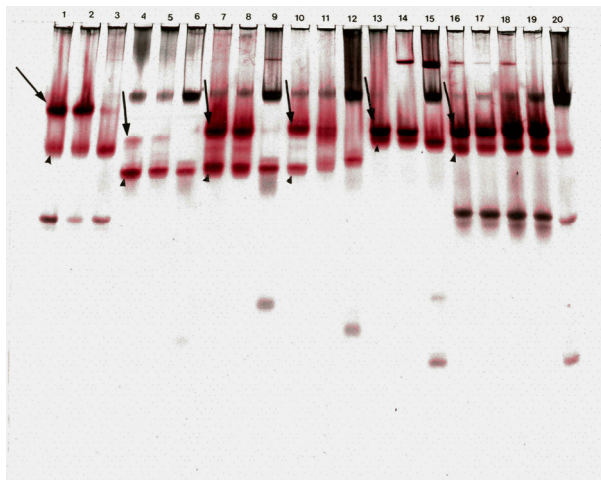


Fig. 1. Esterase pattern in 10% PAGE for late third instar larvae and adult females of *Drosophila mulleri* (1 and 2 = larvae; 3 = female), *D. arizonae* (4 and 5 = larvae; 6 = female), *D. mojavenensis* (7 and 8 = larvae; 9 = female), *D. navojoa* (10 and 11 = larvae; 12 = female), *D. wheeleri* (13 and 14 = larvae; 15 = female), *D. aldrichi* (16 to 19 = larvae; 20 = female). All wells contain individual samples, except for wells 18 and 19, which contain 2 larvae of *D. aldrichi*. Arrow = EST-4; arrowhead = EST-5.

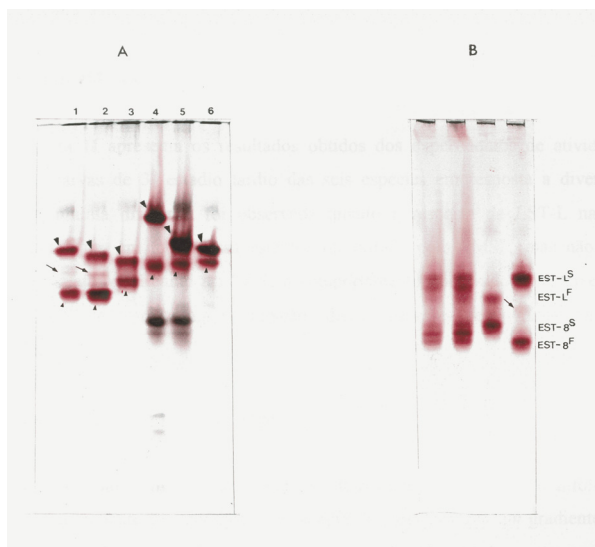


Fig. 2. A. 10% PAGE showing the electrophoretic staining differences for EST-4 and EST-5. 1- *D. mojavenensis*; 2 - *D. arizonae*; 3 - *D. navojoa*; 4 - *D. mulleri*; 5 - *D. aldrichi*; 6 - *D. wheeleri*. B. 10% PAGE of late third instar larvae of *D. navojoa*, showing the different phenotypes observed. Arrows in A and B indicate the interlocus heterodimer EST-4/EST-5. Larger arrowhead = EST-4; smaller arrowhead = EST-5.

Despite the observation of homozygotes for EST-4 in five out of six species analyzed, the quaternary structure for this enzyme as a dimer could be deduced from the presence of a heterodimer between EST-4 and EST-5 in *D. mojavensis* and *D. arizonae* (Figure 2A). This dimeric structure for EST-4 and EST-5 was confirmed by hybrid analyses. In *Drosophila navojoa*, in addition to the presence of the heterodimer, three phenotypes were observed in gels for EST-4 and EST-5 (Figure 2B): homozygous for a slower band (EST-4<sup>S</sup> and EST-5<sup>S</sup>, respectively); homozygous for a faster band (EST-4<sup>F</sup> and EST-5<sup>F</sup>, respectively); and heterozygous, with a three-band pattern. These patterns reinforce the quaternary structure of these enzymes for this species. The same results were observed for EST-5 of *D. mulleri* (data not shown).

### 3.2 Pattern of esterase activity in the presence of inhibitors

Table 2 shows the results obtained for the esterase activity patterns of late third instar larvae of the analyzed species in the presence of different inhibitors. All species showed the same pattern for EST-4: they were inhibited only by PMSF. For EST-5, all species were inhibited only by malathion. No other inhibitor affected the activity of either esterase.

Species	Enzyme	PMSF	Malathion	CuSO <sub>4</sub>	IAC	E-64	Eserine	pCMB	HgCl <sub>2</sub>
<i>D. mulleri</i>	EST-4	++	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	EST-5	⊗	++	⊗	⊗	⊗	⊗	⊗	⊗
<i>D. aldrichi</i>	EST-4	++	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	EST-5	⊗	++	⊗	⊗	⊗	⊗	⊗	⊗
<i>D. wheeleri</i>	EST-4	++	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	EST-5	⊗	++	⊗	⊗	⊗	⊗	⊗	⊗
<i>D. mojavensis</i>	EST-4	++	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	EST-5	⊗	++	⊗	⊗	⊗	⊗	⊗	⊗
<i>D. arizonae</i>	EST-4	++	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	EST-5	⊗	++	⊗	⊗	⊗	⊗	⊗	⊗
<i>D. navojoa</i>	EST-4	++	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	EST-5	⊗	++	⊗	⊗	⊗	⊗	⊗	⊗

Table 2. Esterase activity patterns of EST-4 and EST-5 for the six *Drosophila* species analyzed in the presence of different inhibitors. PMSF = phenylmethylsulfonyl fluoride; Eserine = eserine sulfate; CuSO<sub>4</sub> = copper sulfate; IAC = iodoacetamide; E-64 = trans-epoxysuccinyl-L-leucyl-amido(4-guanidino) butane; pCMB = p-chloromercuribenzoate; HgCl<sub>2</sub> = mercuric chloride. ++ activity inhibited; ⊗ activity not affected.

### 3.3 Isoelectric point (I.P.) determination

The I.P. determination was performed in two phases. In the first phase, we verified that the best range for I.P. determination was 6.0 to 8.0. In the second phase, an ampholyte solution was used for this pH range. Table 3 shows that all esterases presented I.P. between 6.0 and 7.0. As expected, the I.P. values for EST-5 in both larvae and adults of the same species were equal, ranging from 6.47 (*D. navojoa*) to 6.64 (*D. aldrichi*). EST-4 showed a wider range of I.P. variation than EST-5, with *D. mulleri* and *D. navojoa* showing the highest and lowest I.P. values (6.88 and 6.37, respectively).

<i>D. mulleri</i> cluster	I.P.	<i>D. mojavensis</i> cluster	I.P.
<i>D. mulleri</i>		<i>D. mojavensis</i>	
EST-4	6.88	EST-4	6.38
EST-5 (larvae and adult)	6.51	EST-5 (larvae and adult)	6.49
<i>D. aldrichi</i>		<i>D. arizonae</i>	
EST-4	6.55	EST-4	6.53
EST-5 (larvae and adult)	6.64	EST-5 (larvae and adult)	6.56
<i>D. wheeleri</i>		<i>D. navojoa</i>	
EST-4	6.59	EST-4	6.37
EST-5 (larvae and adult)	6.53	EST-5 (larvae and adult)	6.47

Table 3. Isoelectric points for EST-4 and EST-5 of larvae and adults of the six analyzed *Drosophila* species, obtained through the comparison of esterase band mobility in gels with an I.P. marker (hemoglobin; I.P. = 7.1) in a pH range between 6.0 and 8.0.

### 3.4 MW determination

To determine the MW of both enzymes in all six *Drosophila* species analyzed, the technique described by Mateus et al. (2009) was applied using 6% to 12% PAGE and the same MW markers. The results presented there are part of this study. Therefore, in this study, we present the results that were not shown in Mateus et al. (2009), i.e., the MW determinations of EST-4 and EST-5 for *D. mulleri*, *D. aldrichi*, *D. wheeleri* and *D. navojoa*. After electrophoresis, the relative mobility (Rm) values for the esterases of these four species and the molecular markers were obtained. The graphs of Rm versus gel concentration for each MW marker resulted in a different slope. These slopes were plotted against the MW (Figure 1 – Mateus et al., 2009). Ferguson's plot (Log Rm versus gel concentrations) for EST-4 and EST-5 of *D. mulleri*, *D. aldrichi*, *D. wheeleri* and *D. navojoa* are shown in Figure 3.

The plots for both esterases were parallel in all species, indicating that these enzymes have different charges and/or tridimensional structures but very similar molecular weights. From these graphs, the slope was obtained for each enzyme in each species. These values were used to estimate the MW in each case, using the equation  $Y = A + BX$ , where A is the intercept of the Y-axis (2.18766), and B is the slope (0.09452). The slopes and molecular weights are presented in Table 4.

The slope values for both enzymes in all species were similar. EST-5 had more variation, ranging from  $-10.05407 \pm 0.29546$  for *D. navojoa* to  $-11.03429 \pm 0.30178$  for *D. mulleri*. EST-4 was less variable, ranging from  $-10.08361 \pm 0.33581$  for *D. wheeleri* to  $-10.52607 \pm 0.44878$  for *D. mulleri*. The MWs, estimated from these slope values (Table 4), were very close to each other. For EST-4, the MW ranged from 83.537 kD in *D. wheeleri* to 88.218 kD in *D. mulleri*. For EST-5, the MW ranged from 83.225 kD in *D. navojoa* to 93.595 kD in *D. mulleri*. The MWs obtained, including standard deviations, were all approximately 80 kD to 96.8 kD.

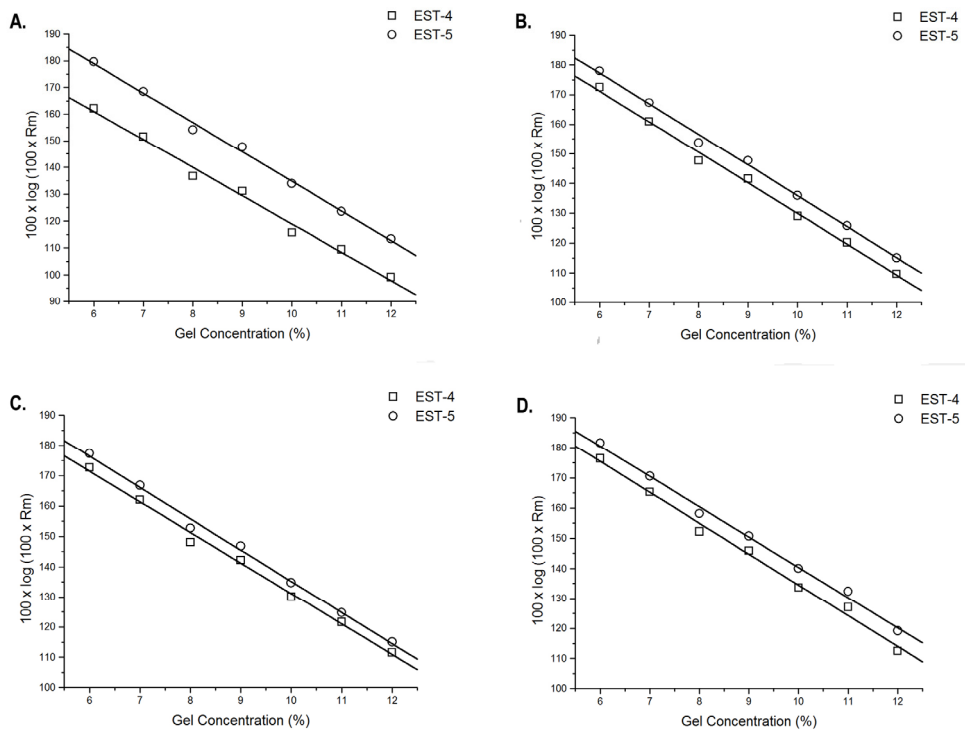


Fig. 3. Log of Rm versus gel concentrations relationship (Ferguson's plot) for EST-4 and EST-5. A - *Drosophila mulleri*; B - *Drosophila aldrichi*; C - *Drosophila wheeleri*; D - *Drosophila navojoa*.

	EST-4		EST-5	
<i>D. mulleri</i> cluster	Slope	M.W.	Slope	M.W.
<i>D. mulleri</i>	$-10.52607 \pm 0.44878$	$88.218 \pm 4.748$	$-11.03429 \pm 0.30178$	$93.595 \pm 3.193$
<i>D. aldrichi</i>	$-10.29768 \pm 0.30168$	$85.802 \pm 3.192$	$-10.34904 \pm 0.28362$	$86.346 \pm 3.000$
<i>D. wheeleri</i>	$-10.08361 \pm 0.33581$	$83.537 \pm 3.553$	$-10.32114 \pm 0.30252$	$86.050 \pm 3.201$
<i>D. mojavensis</i> cluster				
<sup>1</sup> <i>D. mojavensis</i>	$-10.45339 \pm 0.27581$	$87.450 \pm 2.918$	$-10.33036 \pm 0.27437$	$86.148 \pm 2.903$
<sup>1</sup> <i>D. arizonae</i>	$-10.27775 \pm 0.32146$	$85.591 \pm 3.401$	$-10.16554 \pm 0.30570$	$84.404 \pm 3.234$
<i>D. navojoa</i>	$-10.24157 \pm 0.38515$	$85.209 \pm 4.074$	$-10.05407 \pm 0.29546$	$83.225 \pm 3.126$

Table 4. Slopes of Log Rm versus gel concentration relationships and MW estimates for EST-4 and EST-5 of the six cactophilic *Drosophila* species analyzed.<sup>1</sup> Data obtained from Mateus et al. (2009).



### 3.5 Interspecies crosses

#### 3.5.1 Crosses between *D. mulleri* and *D. mojavensis*

The cross between *D. mulleri* and *D. mojavensis* showed asymmetric isolation, with many descendents only in the direction of *D. mulleri* females and *D. mojavensis* males. The reciprocal cross did not produce offspring despite the presence of courtship among couples and eggs in the plate. The hybrid larvae were analyzed in 10% PAGE and showed three-band patterns for EST-4 and EST-5 (Figure 4). For both enzymes, the slower band corresponded to the enzyme encoded by *D. mulleri* and the faster to the enzyme encoded by *D. mojavensis*. The intermediate band represented a hybrid enzyme, indicating that EST-4 and EST-5 are dimeric in both species.

For EST-4, the hybrid intermediate bands were located closer to the band encoded by *D. mulleri*, which could be a result of differences in the I.P. values of these enzymes (Table 3). The same was not observed for EST-5, as this enzyme has nearly the same value for both of these species.

#### 3.5.2 Crosses between *D. navojoa* and *D. mojavensis*

Asymmetric isolation was also observed in the cross between *D. navojoa* and *D. mojavensis*. No offspring were obtained in the direction of *D. mojavensis* females and *D. navojoa* males, despite the fact that courtship between couples and eggs on the plate were observed. The cross between *D. navojoa* females and *D. mojavensis* males was very fertile.

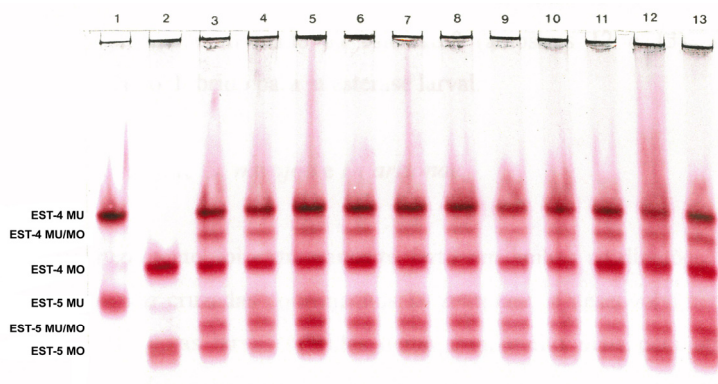


Fig. 4. Esterase pattern in 10% PAGE of late third instar larvae from the parental lines and the hybrids obtained from the cross of *D. mulleri* females and *D. mojavensis* males. 1 = *D. mulleri*; 2 = *D. mojavensis*; 3-13 = hybrid larvae.

The hybrid larvae from this cross were analyzed in 10% PAGE and showed the same three-band patterns observed for *D. mulleri* and *D. mojavensis* (Figure 5). For EST-4, the slower band corresponded to the enzyme encoded by *D. mojavensis*, the faster band corresponded to the enzyme encoded by *D. navojoa*, and the intermediate band corresponded to a hybrid enzyme. The opposite was observed for EST-5: the slower band was from *D. navojoa*, the faster band was from *D. mojavensis*, and an intermediate band was a hybrid enzyme. Again, these results confirm the quaternary structure of both enzymes of these species. An interesting observation was the absence of EST-5 expression in two samples (samples 12 and 13; Figure 5).

### 3.5.3 Crosses between *D. navojoa* and *D. arizonae*

This cross was very fertile in both directions. However, larvae from the cross in the direction of *D. navojoa* females and *D. arizonae* males had very slow development and took much longer to achieve the late third instar stage; they also had a high mortality rate. The larvae analyzed in 10% PAGE from both cross directions presented the same three-band patterns for EST-5. For EST-4, as in both species of the cross, the enzymes had almost the same migration speed under these electrophoretic conditions. One thicker band was observed in the hybrid larvae, which must be the agglomeration of the three bands expected for this enzyme (Figure 6).

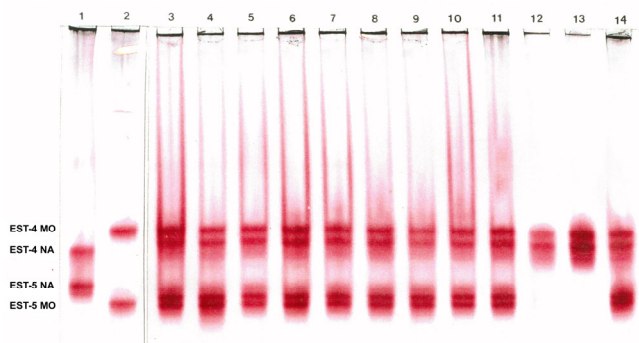


Fig. 5. Esterase pattern in 10 % PAGE of late third instar larvae from the parental lines and the hybrids obtained from the cross of *D. navojoa* females and *D. mojavensis* males. 1 = *D. navojoa*; 2 = *D. mojavensis*; 3-14 = hybrid larvae.

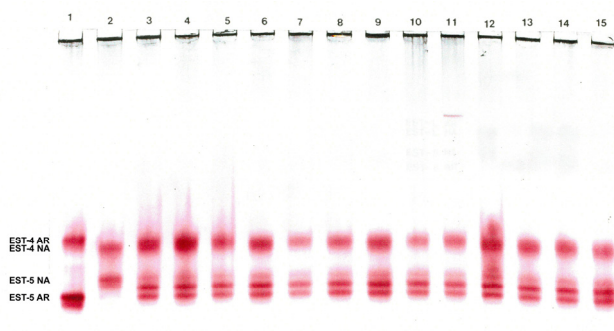


Fig. 6. Esterase pattern in 10% PAGE of late third instar larvae from the parental lines of *D. arizonae* and *D. navojoa* and the hybrids obtained from crosses in both directions. 1 = *D. arizonae*; 2 = *D. navojoa*; 3-15 = hybrid larvae.

### 3.5.4 Crosses between *D. navojoa* and *D. mulleri*

The cross between *D. navojoa* and *D. mulleri* was fertile in both directions. The larvae analyzed by 10% PAGE showed three-band patterns for both EST-4 and EST-5, with the slower enzyme from *D. mulleri* and the faster band from *D. navojoa*. The intermediate band was a hybrid enzyme. These results confirm the dimeric quaternary structure of these

enzymes in these species (Figure 7). Again, the EST-4 hybrid band migrated closer to the slower band from *D. mulleri*, which could be a consequence of the different I.P. values of these enzymes. The same was not observed for EST-5, as they show similar I.P. values for both species.

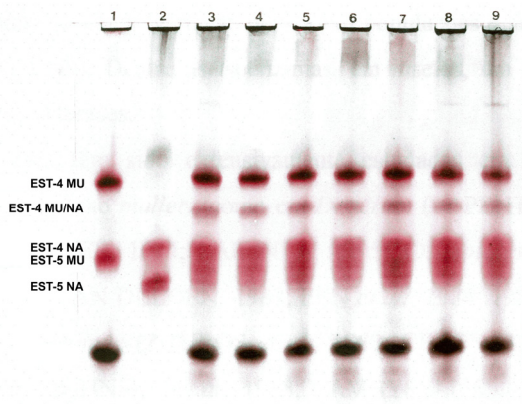


Fig. 7. Esterase pattern in 10% PAGE of late third instar larvae from the parental lines of *D. mulleri* and *D. navojoa* and the hybrids obtained from crosses in both directions. 1 = *D. mulleri*; 2 = *D. navojoa*; 3-9 = hybrid larvae.

#### 4. Discussion

Isozymes are very important in insects and have been used to understand biological problems in several fields of research, including population genetics and systematics, tissue organization, development, metamorphosis, gene regulation and protein synthesis and gene duplication (R. P. Wagner & Selander, 1974). The set of proteins known as esterases constitute one of the most heavily studied groups of isozymes. In the *Drosophila mulleri* complex, which is the subject of this study, esterases have been extensively studied in several species, including *D. serido* (Lapenta et al., 1995, 1998), *D. buzzatii* (East, 1982; Barker, 1994; Lapenta et al., 1995, 1998), *D. mojavensis* (Zouros et al., 1982; Zouros & Van Delden, 1982; Pen et al., 1984, 1986a, 1986b; Mateus et al., 2009), *D. arizonae* (Zouros et al., 1982; Ceron, 1988; Mateus et al., 2009), *D. aldrichi* (F. M. Johnson et al., 1968; Kambysellis et al., 1968) and *D. mulleri* (F. M. Johnson et al., 1968; Kambysellis et al., 1968; Richardson & Smouse, 1976; Ceron, 1988).

Zouros et al. (1982) detected two esterases with different patterns of temporal and tissue-specific expression in *Drosophila mojavensis* and *D. arizonae* (formerly *D. arizonensis*). They detected a specific  $\beta$ -esterase of the late third instar phase of larval development and in the carcass, named EST-4, in contrast to another  $\beta$ -esterase, named EST-5, which is expressed during all developmental phases and is found predominantly in hemolymph and the fat body. They proposed that the most likely hypothesis is that both enzymes are products of a gene duplication that occurred prior to the speciation of the *D. repleta* group, and their patterns of tissue-specific and temporal expression diverged more recently. This hypothesis was suggested because the enzymes show interlocus heterodimers, different patterns of expression (Zouros et al., 1982) and 82% identity in their N-terminal amino acid sequences

(Pen et al., 1986a; Pen et al., 1990). More recently, Robin et al. (2009) proposed that, in *D. mojavensis*, these enzymes are most likely encoded by two genes, *Est-2a* (EST-5) and *Est-2c* (EST-4), which are products of one gene duplication out of a total of eleven duplications that explain the evolution of the catalytic  $\beta$ -esterase cluster in the *Drosophila* genus (five in the *Sophophora* and 6 in the *Drosophila* subgenus).

Our results reinforce the hypothesis proposed by Zouros et al. (1982), extending the knowledge about these enzymes as products of a gene duplication to other *D. mulleri* complex species. All six analyzed species show distinct temporal expression patterns for EST-4 and EST-5, with EST-4 showing activity only at the end of the third instar larval stage (Figure 1). The inhibition experiments (Table 2) showed that EST-4 has the same pattern for all six species: it is inhibited by PMSF and not affected by malathion. The opposite was observed for EST-5 in all six species: it was inhibited by malathion and not affected by PMSF. The other inhibitors tested (eserine sulfate, copper sulfate, iodoacetamide and E-64) had no effect on the activity of either enzyme. Moreover, the presence of homozygotes and heterozygotes for EST-5 independent of the EST-4 genotype in *D. navojoa* (Figure 2) and *D. mulleri* (data not shown) support the idea of an independent origin of these enzymes from two distinct loci. Despite these differences, these enzymes display similar features, such as structural similarities (Pen et al., 1986a; Pen et al. 1990) that allow the formation of dimers in *D. mojavensis*, *D. arizonae* and *D. navojoa* (Figure 2).

The gene duplication process is considered one of the most important mechanisms of the generation of new genes and functions during the evolutionary process. Jeffreys & Harris (1982) suggested gene duplication mechanisms that could happen to genes during evolution. Among the mechanisms presented, the most likely mechanism that could have generated the EST-4 and EST-5 loci is the same mechanism that might have generated the globin family, that is, *in tandem* gene duplication by pairing errors during meiosis that cause unequal crossing-over because of the presence of short repeat sequences located in the 3' and 5' ends of the unduplicated ancestral gene.

The EST-5 gene in *D. pseudoobscura* is a good example of gene duplication with later divergence (Brady & Richmond, 1992). The EST-5 enzyme is encoded by the *Est-5B* gene, which is expressed during the life cycles of all insects and is linked to two other genes, *Est-5A* and *Est-5C*, on the X chromosome (Brady et al., 1990). In *D. melanogaster*, the homologous gene is *Est-6*, which codes for the enzyme EST-6 during the insect's life cycle and has only one grouped gene, *Est-P* (Collet et al., 1990). Both *Est-5A* of *D. pseudoobscura* and *Est-P* of *D. melanogaster* are expressed only at the third instar larval stage, producing only one transcript. On the other hand, *Est-5C* of *D. pseudoobscura* is not expressed in any developmental phase (Brady et al., 1990). According to Brady & Richmond (1992), who compared the DNA sequences of coding and flanking regions of all three *D. pseudoobscura* and two *D. melanogaster* genes, only two genes, which are already products of a gene duplication, were present before these two species diverged. These two ancestral genes were probably *Est-5A* and *Est-5B* in the first species and *Est-6* and *Est-P* in the second species. A second duplication occurred later in *D. pseudoobscura*, giving rise to the *Est-5C* gene. However, the findings of Robin et al. (2009) contrast with the evolutionary model proposed by Brady & Richmond (1992); in their analyses, the *Est-5A/Est-5B* duplication (which they call *Est6/7*) occurred after the *melanogaster/obscura* group divergence, whereas Brady & Richmond (1992) place this duplication prior to the divergence.

In our case, Zouros et al. (1982) proposed that the genes coding for EST-4 and EST-5 (*Est-2c* and *Est-2a*, respectively, according to Robin et al., 2009) were also products of a duplication event prior to the divergence of the species that belong to the *D. repleta* group and that the EST-4 gene was later inactivated in some species of this group, including *D. tira*, *D. hydei* and *D. eohydei*. Moreover, the lower activity of EST-4 in *D. mulleri*, *D. aldrichi*, *D. repleta* and *D. peninsularis* could indicate this EST-4 inactivation process. However, our results showed that *D. mulleri*, *D. aldrichi* and *D. wheeleri* had high EST-4 activities compared to the other species (Figure 2A). This difference in the level of activity of EST-4 for the same species in these studies could be the result of differences in the origins of the lines used in each study. Therefore, the populations of *D. mulleri* and *D. aldrichi* that were analyzed by Zouros et al. (1982) could have a certain degree of EST-4 inactivation that was not observed in the present study.

The enzymes analyzed in this study had biochemical differences compared to other esterases of other *Drosophila* species. For example, the I.P. values for EST-4 and EST-5 for the six *Drosophila* species analyzed were between 6.0 and 7.0 (Table 3). These values were different from those of *D. melanogaster* obtained by Healy et al. (1991), as only 2 out of 15 esterases had I.P. values close to the values obtained in this study (between 6.0 and 7.0). All others showed values below 6.0, with the majority of values between 4.0 and 5.0.

Regarding the MWs of these enzymes, our results are in agreement with previous studies that estimated this parameter. EST-4 had MW values between 83 and 89 kD, and EST-5 had MW values between 83 and 94 kD (Table 4), which are very close to the MWs obtained by Pen et al. (1984), which were between 85 and 95 kD for a variant of the EST-4 (with altered specificity to  $\alpha$ -naphthyl acetate) using gel filtration chromatography. Pen et al. (1984) also used denaturing gel electrophoresis (SDS-PAGE) and obtained the MWs of the subunits of EST-4 as 62–64 kD. In another study, Pen et al. (1986a) determined the MWs for the subunits of EST-5 as 64–66 kD. Regardless of the method used and the different results obtained (for the entire protein or for subunits), EST-4 had a smaller MW than EST-5, as observed in this study.

The interspecies crosses performed in this study had results that were in accordance with the known phylogenetic relationships among the species analyzed. This information is based on the morphological work of Throckmorton (1982) and Vilela (1983), the cytological work of Wasserman (1982, 1992 for reviews), several allozyme studies (Zouros, 1973; Richardson et al., 1975; Richardson and Smouse, 1976; Richardson et al., 1977; Heed et al., 1990), molecular studies (Sullivan et al., 1990; Russo et al., 1995; Spicer, 1995, 1996) and an analysis using multiple sources of characters (Durando et al., 2000). The crosses between *D. mulleri* and *D. mojavensis* showed the same results of those of Patterson & Crow (1940) and Bicudo (1982), with offspring obtained only in the direction of *D. mulleri* females and *D. mojavensis* males. For *D. navojoa* crossed with *D. mojavensis*, an F1 was produced only in the direction of *D. navojoa* females and *D. mojavensis* males. In this case, Ruiz et al. (1990) observed descendants in both directions but a very low percentage of offspring, depending on the geographic lineage used, in the direction in which we detected isolation. In crosses between *D. navojoa* and *D. arizonae*, both directions were fertile, which was also found by Ruiz et al. (1990). Finally, in crosses between *D. navojoa* and *D. mulleri*, we detected descendants in both directions, in contrast to the results of Bicudo (1982), who found fertility only in the direction of *D. mulleri* females and *D. navojoa* males.

In all of these crosses, the phenotypic observations of the esterase patterns from late third instar hybrid larvae produced three bands for both EST-4 and EST-5 (Figures 4, 5, 6 and 7), except for larvae from the cross between *D. navojoa* and *D. arizonae*, which produced a thicker band because the parental bands have almost the same migration pattern under the electrophoretic conditions used in this study. These results indicate that in all six *Drosophila* species, EST-4 and EST-5 have dimeric quaternary structures. Another important observation from some of these crosses was the presence of hybrid larvae with no EST-5 activity (*D. navojoa* x *D. mojavensis* – Figure 5; *D. navojoa* x *D. arizonae* – data not shown). These results indicate that some hybrid larvae had problems with the regulation of *Est-2a* gene expression, which most likely codes for the EST-5 enzyme, without affecting the expression of its homologous gene, *Est-2c*, which most likely codes for the EST-4 enzyme (Robin et al., 2009). These results reinforce the idea that these two loci are independent.

The possible role of EST-4 in these *Drosophila* species remains an open question. According to Holmes & Masters (1967, as cited in Oakeshott et al., 1993), esterases can be classified into four types through their specific inhibition patterns. Carboxylesterases are esterases that are inhibited only by organophosphates, such as paraoxon, fenitrooxon and DFP (diisopropylfluorophosphate). Cholinesterases are inhibited by organophosphates and carbamates, such as eserine sulfate. Arylesterases are inhibited only by sulfhydrylic agents, such as p-chloromercuribenzoate (pCMB). Acylesterases are not inhibited by any of these agents. Inhibition of EST-5 only by malathion, an organophosphate, suggests that this enzyme belongs to the class of carboxylesterases. Inhibition of EST-4 by PMSF and the absence of inhibition in the presence of all other inhibitors tested suggest that this enzyme probably belongs to the class of acylesterases.

According to Augustinsson (1968), esterases are closely related to the class of serine-proteases, forming a multigenic family of serine-hydrolases. The main features that support this hypothesis are the three consensus amino acid residues that are present in the active site of esterases and serine-proteases, including an invariant serine, enzymatic inactivation by DFP, which binds irreversibly to the serine residue of both enzymes, inhibition by organophosphates and carbamates and the superposition of substrate preference (Augusteyn et al., 1969; Krisch, 1971; Dayhoff et al., 1972; Heymann, 1980; Previero et al., 1983; as cited in Myers et al., 1988). However, Myers et al. (1988) showed that some esterases cannot be included in this multigenic family because they do not have the same amino acid residues in the charge exchange system of the enzyme active site.

The absence of EST-4 and Est-5 inhibition by copper sulfate and iodoacetamide, combined with data for E-64, which is a diagnostic inhibitor of cysteine-proteases, indicate that neither enzyme has an essential cysteine residue in its active site. On the other hand, the inhibition of EST-4 by PMSF, which is a diagnostic compound for serine-proteases and other enzymes with a serine residue in the active site, and of EST-5 only by malathion indicated that both enzymes have an important serine residue in the active site, suggesting that they belong to the class of serine-hydrolases. As these enzymes display high esterase activity, we can postulate that they are serine-esterases (Holmes & Masters, 1967). The multigenic family of serine-esterases includes several enzymes with a wide range of functions, including cholinesterases, lipases, lysophospholipases, cholesterol-esterases, non-specific carboxylesterases and juvenile hormone esterases (Ryger et al., 1989; Doctor et al., 1990; Shimada et al., 1990; as cited in Myers et al., 1993). Therefore, EST-4 and EST-5 probably belong to this multigenic family, with EST-4 as an acylesterase (E.C. 3.1.1.6) and EST-5 as a non-specific carboxylesterase (E.C. 3.1.1.1).

To establish the possible role of EST-4, the following information must be considered. Healy et al. (1991) observed that all *D. melanogaster* acetyl esterases are inhibited by OTFP (3-octylthio-1,1,1-trifluoropropan-2-one), which is a powerful inhibitor of juvenile hormone esterase activity in Lepidoptera, suggesting that all acetyl esterases from this species have similar properties as juvenile hormone esterase. Moreover, East (1982) proposed that esterase-J from *D. buzzatii*, which is supposedly the enzyme from this species that corresponds to EST-4 in this study, is a juvenile hormone esterase, acting together with EST-1 in the larval phase to control the levels of this hormone. In the adult phase, only EST-1 would be responsible for this control. On the other hand, EST-2 could be the enzyme responsible for digestive and detoxification processes and ester absorption in adults. EST-4 has a very tissue-specific and temporal pattern of expression, which indicates that there is a specific regulatory system that controls its expression at a specific tissue (carcass) and period of time (at the end of the larval phase, when all of the processes for pupation have been initiated). Therefore, as an acetyl esterase with a very specific temporal expression pattern, EST-4 could be involved in these transformation processes, acting as an auxiliary enzyme for the degradation of juvenile hormone esterase. The degradation of this hormone in this phase allows the liberation of prothoracicotropic hormone by the brain, which stimulates ecdysone production by the prothoracic gland, initiating metamorphosis (Coundron et al., 1981). However, analyzing the EST-4 inhibition data alone could lead to the hypothesis that this enzyme could be a serine-protease that also has esterase activity and is involved in a proteolytic activity during the larva-pupae conversion process; it is likely to be involved in this process. Regarding EST-5, considering the fact that it is expressed during the entire life cycle of the insect and is found mainly in the hemolymph and fat body, it is a non-specific carboxylesterase that is probably involved in digestive processes.

## 5. Conclusions

This study contributes to a better understanding of the differentiation of two enzymes that are products of a gene duplication in six cactophilic *Drosophila* species. We present additional evidence to support the gene duplication event that gave rise to the genes responsible for the EST-4 and EST-5 enzymes, which are the main  $\beta$ -esterases found in several species of the *D. mulleri* complex of the *D. repleta* group. We also contribute to the elucidation of the possible physiological roles of these esterases in this group. Further steps in this investigation will be to determine specific biochemical parameters of both enzymes after purification. We are also interested in identifying the changes that occur in the regulatory system of gene expression that lead to differentiation in the patterns of tissue-specific and temporal expression of these enzymes; that is, understanding what triggers EST-4 expression only in the late third instar larvae and at the larval carcass. We are also interested in determining the intra- and/or extracellular processes in which these enzymes are involved and their interacting molecules. Thus, we will be able to complement this initial step with an increased understanding of the differentiation of these two genes that result from a gene duplication event.

## 6. Acknowledgments

We would like to thank CNPq for funding Rogério P. Mateus (Master's degree fellowship). We would also like to thank CAPES, FINEP and FUNDUNESP for supporting this work,

Eliani Nobuco Ikeguchi Ohira for technical support and Dr. Hermione Elly Melara Campos Bicudo for *Drosophila* stocks.

## 7. References

- Argentine, J.A., & James, A.A. (1995). Characterization of a salivary gland-specific esterase in the vector mosquito, *Aedes aegypti*. *Insect Biochem. Molec. Biol.*, Vol. 25, No. 5, pp. 621-630
- Augustinsson, K.B. (1968). The evolution of esterases in vertebrates, In: *Homologous enzymes and biochemical evolution*, N.V. Their & J. Roche, pp. 299-311, Gordon & Breach, New York
- Barker, J.S.F. (1994). Sequential gel electrophoretic analysis of esterase-2 in two populations of *Drosophila buzzatii*. *Genetica*, Vol. 92, No. 3, pp. 165-175
- Bicudo, H.E.M.C. (1982). *Estudo citogenético de espécies de Drosophila do complexo mulleri (grupo repleta): A regulação da atividade organizadora nucleolar*. Tese de Livre Docência, IBILCE/UNESP, São José do Rio Preto
- Brady, J.P., & Richmond, R.C. (1992). An evolutionary model for the duplication and divergence of esterase genes in *Drosophila*. *J. Mol. Evol.*, Vol. 34, pp. 506-521
- Brady, J.P., Richmond, R.C., & Oakeshott, J.G. (1990). Cloning of the esterase-5 locus from *Drosophila pseudoobscura* and comparison with its homologue in *D. melanogaster*. *Mol. Biol. Evol.*, Vol. 7, pp. 525-546
- Bridges, C.B. (1936). The Bar "gene" duplication. *Science*, Vol. 83, pp. 210-211
- Ceron, C.R. (1988). *Padrão de esterases no desenvolvimento de Drosophila mulleri, D. arizonae e seus híbridos*. Tese de Doutorado, Departamento de Biologia, IB-USP, São Paulo
- Ceron, C.R., Santos, J.R., & Bicudo, H.E.M.C. (1992). The use of gelatin to dry cellophane wound slab gels in an embroidering hoop. *Braz. J. Genet.*, Vol. 15, No. 1, pp. 201-203
- Collet, C., Nielsen, K.M., Russell, R.J., Karl, M., Oakeshott, J.G., & Richmond, R.C. (1990). Molecular analysis of duplicated esterase genes in *Drosophila melanogaster*. *Mol. Biol. Evol.*, Vol. 7, pp. 9-28
- Coundron, T.A., Dunn, P.E., Seballos, H.L., Wharen, R.E., Sanburg, L.L., & Law, J.H. (1981). Preparation of homogeneous juvenile hormone specific esterase from the haemolymph of the tobacco hornworm, *Manduca sexta*. *Insect Biochem.*, Vol. 11, No. 4, pp. 453-461
- Davis, B.J. (1964). Disc electrophoresis. II. Methods and application to human serum proteins. *Annals N. Y. Acad. Sci.*, Vol. 121, pp. 404-427
- Durando, C.M., Baker, R.H., Etges, W.J., Heed, W.B., Wasserman, M., & DeSalle, R. (2000). Phylogenetic analysis of the *repleta* species group of the genus *Drosophila* using multiple sources of characters. *Molecular Phylogenetics and Evolution*, Vol. 16, pp. 296-307
- East, P.D. (1982). Non-specific esterases of *Drosophila buzzatii*, In: *Ecological genetics and evolution. The cactus-yeast-Drosophila model system*, J.S.F. Baker & W.T. Starmer, pp. 323-338, Academic Press, Australia
- Feyereisen, R. (1995). Molecular biology of insecticide resistance. *Toxicol. Lett.*, Vol. 82/83, pp. 83-90
- Fogleman, J.C., & Abril, J.R. (1990). Ecological and evolutionary importance of host plant chemistry, In: *Ecological and Evolutionary Genetics of Drosophila*, J.S.F. Barker, W.T. Starmer, & R. MacIntyre, pp. 121-143, Plenum Press, New York



- Fogleman, J.C., & Danielson, P.B. (2001). Chemical interactions in the cactus-microorganism-*Drosophila* model system of the Sonoran desert. *Am. Zool.*, Vol. 41, pp. 877-889
- Gottlieb, L.D. (1982). Conservation and duplication of isozymes in plants. *Science*, Vol. 216, pp. 373-380
- Gu, X., & Zera, A.J. (1994). Developmental profiles and characteristics of hemolymph juvenile hormone esterase, general esterase and juvenile hormone binding in the cricket, *Gryllus assimilis*. *Comp. Biochem. Physiol. B Comp. Biochem. Mol. Biol.*, Vol. 107, No. 4, pp. 553-560
- Hart, G.E. (1983). Genetics and evolution of multilocus isozymes in hexaploid wheat. *Isozymes Curr. Top. Biol. Med. Res.*, Vol. 10, pp. 365-380
- Healy, M.J., Dumancic, M.M., & Oakeshott, J.G. (1991). Biochemical and physiological studies of soluble esterases from *Drosophila melanogaster*. *Biochem. Genet.*, Vol. 29, pp. 365-388
- Hedrick, J.L., & Smith, A.J. (1968). Size and charge isomer separation and estimation of molecular weights of proteins by disc gel electrophoresis. *Arch. Biochem. Biophys.*, Vol. 126, pp. 155-164
- Heed, W.B., Sanchez, A., Armengol, R., & Fontdevila, A. (1990). Genetic differentiation among island populations and species of cactophilic *Drosophila* in the West Indies, In: *Ecological and Evolutionary Genetics of Drosophila*, J.S.F. Barker, W.T. Starmer, & R. J. MacIntyre, pp. 447-490, Plenum Press, New York
- Holmes, R.S., Masters, C.J. (1967). The developmental multiplicity and isoenzyme status of cavian esterases. *Biochim. Biophys. Acta*, Vol. 132, pp. 379-399
- Jeffreys, A.J., Harris, S. (1982). Processes of gene duplication. *Nature*, Vol. 296, pp. 9-10
- Johnson, G.B. (1973). Importance of substrate variability to enzyme polymorphism. *Nature New Biol.*, Vol. 243, pp. 151-153
- Johnson, G.B. (1974). Enzyme polymorphism and metabolism. *Science*, Vol. 184, pp. 28-37
- Johnson, F.M., Richardson, R.H., & Kambyzellis, M.P. (1968). Isozyme variability in species of the genus *Drosophila*. III. Qualitative comparison of esterases of *D. aldrichi* and *D. mulleri*. *Biochem. Genet.*, Vol. 1, pp. 239-247
- Kambyzellis, M.P., Johnson, F.M., & Richardson, R.H. (1968). Isozyme variability in species of the genus *Drosophila*. IV. Distribution of the esterases in the body tissues of *D. aldrichi* and *D. mulleri* adults. *Biochem. Genet.*, Vol. 1, pp. 249-265
- Karotam, J., Delves, A.C., & Oakeshott, J.G. (1993). Conservation and change in structural and 5' flanking sequences of esterase 6 in sibling *Drosophila* species. *Genetica*, Vol. 88, pp. 11-28
- Kircher, H.W. (1982). Chemical composition of cacti and its relationship to Sonoran desert *Drosophila*, In: *Ecological genetics and evolution: the cactus-yeast-Drosophila model system*, J.S.F. Barker & W.T. Starmer, pp. 143-158, Academic Press, Sydney, Australia
- Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., & Koonin, E.V. (2002). Selection in the evolution of gene duplications. *Genome Biology*, Vol. 3, pp. 0008.1-0008.9
- Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, Vol. 227, pp. 680-685
- Lapenta, A.S., Bicudo, H.E.M.C., Ceron, C.R., & Manzato, J.A. (1995). Esterase patterns of species in the *Drosophila buzzatii* cluster. *Cytobios*, Vol. 84, pp. 13-29

- Lapenta, A.S., Bicudo, H.E.M.C., Ceron, C.R., & Cordeiro, J.A. (1998). Esterase patterns and phylogenetic relationships of species and strains included in the *Drosophila buzzatii* cluster. *Cytobios*, Vol. 96, pp. 95-107
- Lewis, E.B. (1951). Pseudoallelism and gene evolution. *Cold Spring Harb. Symp. Quant. Biol.*, Vol. 16, pp. 159-174
- Manfrin, M.H., & Sene, F.M. (2006). Cactophilic *Drosophila* in South America: a model for evolutionary studies. *Genetica*, Vol. 126, pp. 57-75
- Markow, T. (1981). Courtship behavior and control of reproductive isolation between *Drosophila mojavensis* and *Drosophila arizonensis*. *Evolution*, Vol. 35, No. 5, pp. 1022-1026
- Mateus, R.P., & Sene, F.M. (2007). Population genetic study of allozyme variation in natural populations of *Drosophila antonietae* (Insecta, Diptera). *Journal of Zoological Systematics and Evolutionary Research*, Vol. 45, pp. 136-143
- Mateus, R.P., Cabral, H., Bonilla-Rodriguez, G.O., & Ceron, C.R. (2009) Molecular weight estimation of esterase isoenzymes in closely related *Drosophila* species (Diptera: Drosophilidae) in non-denaturing polyacrylamide gel electrophoresis. *Braz. Arch. Biol. Technol.*, Vol. 52, pp. 1083-1089
- Matzkin, L.M. (2005). Activity variation in alcohol dehydrogenase paralogs is associated with adaptation to cactus host use in cactophilic *Drosophila*. *Mol. Ecol.*, Vol. 14, pp. 2223-2231
- Matzkin, L.M., Watts, H.D., Bitler, B.G., Machado, C.A., & Markow, T.A. (2006). Functional genomics of cactus host shifts in *Drosophila mojavensis*. *Mol. Ecol.*, Vol. 15, pp. 4635-4643
- Morais, P.B., Rosa, C.A., Hagler, A.N., Mendonça-Hagler, L.C. (1994). Yeast communities of the cactus *Pilosocereus arrabidaei* as resources for larval and adult stages of *Drosophila serido*. *Antonie van Leeuwenhoek*, Vol. 66, pp. 313-317
- Myers, M.A., Richmond, R.C., & Oakeshott, J.G. (1988). On the origins of esterases. *Mol. Biol. Evol.*, Vol. 5, No. 2, pp. 113-119
- Nei, M. (1969). Gene duplication and nucleotide substitution in evolution. *Nature*, Vol. 221, pp. 40-42
- Oakeshott, J.G., van Papenrecht, E.A., Boyce, T.M., Healy, M.J., & Russel, R.J. (1993). Evolutionary genetics of *Drosophila* esterases. *Genetica*, Vol. 90, pp. 239-268
- Ohno, S. (1970). *Evolution by gene duplication*, Springer-Verlag, New York
- Patterson, J.T., & Crow, J.F. XII. Hybridization in the mulleri group of *Drosophila*. *Univ. Texas Publ.*, Vol. 4032, pp. 251-256
- Pen, J., Rongen, H.A.H., & Beintema, J.J. (1984). Purification and properties of esterase-4 form *Drosophila mojavensis*. *Biochem. Biophys. Acta*, Vol. 789, pp. 203-209
- Pen, J., van Beeumen, J., & Beintema, J.J. (1986a). Structural comparison of two esterases from *Drosophila mojavensis* isolated by immunoaffinity chromatography. *Biochem. J.*, Vol. 238, pp. 691-699
- Pen, J., Schipper, A., Rongen, H.A.H., & Beintema, J.J. (1986b). Differences in specificity and catalytic efficiency between allozymes of esterase-4 from *Drosophila mojavensis*. *Mol. Biol. Evol.*, Vol. 3, No. 4, pp. 366-373
- Pen, J., Bolks, G.J., Hoeksema-Du Pui, M.L.L., & Beintema, J.J. (1990). Serine esterases: structural conservation during animal evolution and variability in enzymic properties in the genus *Drosophila*. *Genetica*, Vol. 81, pp. 125-131

- Pereira, M.A.Q.R., Vilela, C.R., Sene, F.M. (1983). Notes on breeding and feeding sites on some species of the *repleta* group of the genus *Drosophila* (Diptera, Drosophilidae). *Ciência e Cultura*, Vol. 35, pp. 1313-1319
- Richardson, R.H., & Smouse, P.E. (1976). Patterns of molecular variation. I. Interspecific comparison of electromorphs in the *Drosophila mulleri* complex. *Biochem. Genet.*, Vol. 14, pp. 447-466
- Richardson, R.H., Richardson, M.E., & Smouse, P.E. (1975). Evolution of electrophoretic mobility in the *Drosophila mulleri* complex, In: *Isozymes IV: Genetics and Evolution*, C.L. Markert, pp. 533-545, Academic Press, New York
- Richardson, R.H., Smouse, P.E., & Richardson, M.E. (1977). Patterns of molecular variation. II. Associations of electrophoretic mobility and larval substrate within species of the *Drosophila mulleri* complex. *Genetics*, Vol. 85, pp. 141-154
- Robin, C., Bardsley, L.M.J., Coppin, C., & Oakeshott, J.G. (2009). Birth and death of genes and functions in the  $\beta$ -esterase cluster of *Drosophila*. *J. Mol. Evol.*, Vol. 69, pp. 10-21
- Ruiz, A., Heed, W.B., & Wasserman, M. (1990). Evolution of the *mojavensis* cluster of cactophilic *Drosophila* with descriptions of two new species. *J. Hered.*, Vol. 81, pp. 30-42
- Russo, C.A.M., Takezaki, N., & Nei, M. (1995). Molecular phylogeny and divergence times of *Drosophilid* species. *Mol. Biol. Evol.*, Vol. 12, pp. 391-404
- Spicer, G.S. (1995). Phylogenetic utility of the mitochondrial cytochrome oxidase gene: Molecular evolution of the *Drosophila buzzatii* species complex. *J. Mol. Evol.*, Vol. 41, pp. 749-759
- Spicer, G.S., & Pitnick, S. (1996). Molecular systematics of the *Drosophila hydei* subgroup as inferred from mitochondrial DNA sequences. *J. Mol. Evol.*, Vol. 43, pp. 281-286
- Starmer, W.T., & Gilbert, D.G. (1982). A quick and reliable method for sterilizing eggs. *Drosophila Information Service*, Vol. 58, pp. 170-171
- Starmer, W.T., Barker, J.S.F., Phaff, H.J., & Fogleman, J.C. (1986). Adaptations of *Drosophila* and yeasts – their interactions with the volatile 2-propanol in the cactus microorganism *Drosophila* model system. *Australian Journal of Biological Sciences*, Vol. 39, pp. 69-77
- Stephens, S.G. (1951). Possible significance of duplication in evolution. *Adv. Genet.*, Vol. 4, pp. 247-265
- Sullivan, D.T., Atkinson, P.W., Bayer, C.A., & Menotti-Raymond, M. (1990). The evolution of *Adh* expression in the *repleta* group of *Drosophila*, In: *Ecological and Evolutionary Genetics of Drosophila*, J.S.F. Barker, W.T. Starmer, & R.J. MacIntyre, pp. 407-418, Plenum Press, New York
- Tidon-Sklorz, R., & Sene, F. M. (1995). Evolution of the *buzzatii* cluster (*Drosophila repleta* group) in the Northeastern South America. *Evolución Biológica*, Vol. 8/9, pp. 71-85
- Throckmorton, L.H. (1982). Pathways of evolution in the genus *Drosophila* and the founding of the *repleta* group, In: *Ecological Genetics and Evolution: The Cactus-Yeast-Drosophila Model System*, J.S.F. Barker & W.T. Starmer, pp. 33-48, Academic Press, New York
- Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R., Zhang, H., & The FlyBase Consortium. (2009). FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Research*, Vol. 37, pp. D555-D559

- Via, S. (1990). Ecological genetics and host adaptation in herbivorous insects: the experimental study of evolution in natural and agricultural systems. *Annu. Rev. Entomol.*, Vol. 35, pp. 421-446
- Vilela, C.R. (1983). A revision of the *Drosophila repleta* species group (Diptera, Drosophilidae). *Rev. Brasil. Entomol.*, Vol. 27, pp. 1-114
- Vogt, R.G., & Riddiford, L.M. (1981). Pheromone binding and inactivation by moth antennae. *Nature*, Vol. 293, pp. 161-163
- Wagner, A. (2002). Selection and gene duplication: a view from the genome. *Genome Biology*, Vol. 3, pp. 1012.1-1012.3
- Wagner, R.P., & Selander, R.K. (1974). Isozymes in insects and their significance. *Annual Review of Entomology*, Vol. 19, pp. 117-138
- Walker, C.H., & Mackness, M.I. (1983). Esterases: problems of identification and classification. *Biochem. Pharmacol.*, Vol. 32, pp. 3265-3269
- Wasserman, M. (1982). Cytological evolution in the *Drosophila repleta* species group, In: *Ecological Genetics and Evolution: The Cactus-Yeast-Drosophila Model System*, J.S.F. Barker & W.T. Starmer, pp. 49-64, Academic Press, New York
- Wasserman, M. (1992). Cytological evolution of the *Drosophila repleta* species group. In: *Drosophila Inversion Polymorphism*, C.B. Krimbas & J.R. Powell, pp. 455-552, CRC Press, Boca Raton, FL
- Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, Vol. 18, pp. 292-298
- Zouros, E. (1973). Genetic differentiation associated with the early stages of speciation in the *mulleri* subgroup of *Drosophila*. *Evolution*, Vol. 27, pp. 601-621
- Zouros, E., van Delden, W., Odense, R., van Dijk, H. (1982). An esterase duplication in *Drosophila*: differences in expression of duplicate loci within and among related species. *Biochem. Genet.*, Vol. 20, pp. 929-942

# SNCA Gene Multiplication: A Model Mechanism of Parkinson Disease

Kenya Nishioka<sup>1</sup>, Owen A. Ross<sup>2</sup> and Nobutaka Hattori<sup>1</sup>

<sup>1</sup>*Department of Neurology, Juntendo University School of Medicine, Tokyo*

<sup>2</sup>*Division of Neurogenetics, Department of Neuroscience, Mayo Clinic Jacksonville FL*

<sup>1</sup>*Japan*

<sup>2</sup>*USA*

## 1. Introduction

Parkinson disease (PD) is caused by several pathogenic mutations in genes such as alpha-synuclein (*SNCA*; MIM#163890), Leucine-rich repeat kinase 2 (*LRRK2*; MIM#609007), *PARKIN* (*parkin*; MIM#602544), PTEN-induced kinase 1 (*PINK1*; MIM#608309), and *DJ-1* (MIM#602533) (Farrer, 2006). The alpha-synuclein protein is also a major component of Lewy bodies (LB), the pathologic substrate that is observed in PD patients at autopsy (Spillantini et al., 1997). LB are generally localized to the mid-brain in patients with PD, however a widespread distribution of LB, including cortical regions, is seen in dementia with Lewy bodies (DLB) (Braak et al., 2003, McKeith et al., 2005). The observation of *SNCA* multiplications co-segregating with PD and dementia in families led to the hypothesis that over-expression of the alpha-synuclein protein is an important mechanism of disease. Herein, we place the gene dosage effect of *SNCA* in PD in perspective and describe the recent molecular insights underlying them.

## 2. *SNCA* triplication and duplication in hereditary PD

PD is the second most frequent neurodegenerative disorder following Alzheimer disease in the elderly. The main symptoms of PD are tremor, bradykinesia, and gait disturbance. PD genetics is categorized into two groups; one is sporadic PD and the other is familial PD. Familial PD has two forms; autosomal dominant heredity (ADPD) and autosomal recessive heredity (ARPD). ADPD has been observed to be caused by mutations in *SNCA* and *LRRK2*. ARPD is caused by homozygous or compound heterozygous mutations in *PARKIN*, *PINK1*, and *DJ-1* (Farrer, 2006). This review will focus on *SNCA* which is located on chromosome 4q21-22 and encodes the 140 amino acid alpha-synuclein protein. *SNCA* has three point mutations; c.88G>C (Ala30Pro), c.188G>A (Glu46Lys) and c.209G>A (Ala53Thr) (Kruger et al., 1998, Zarranz et al., 2004), but they are very rare.

*SNCA* duplications and triplications have also been identified as a genetic cause of ADPD. Duplication has two *SNCA* copies on one allele (50% dose increase) and triplication has three, 100 percent dose increase (Figure 1). Rarely, compound heterozygote forms (two duplication alleles) are seen as *SNCA* triplication events (Ikeuchi et al., 2008). These multiplications generate higher *SNCA* expression of mRNA and protein, the so called gene

dosage effect. Increasing the levels of protein appears to influence the clinical manifestations of PD patients. A subtle increase in alpha-synuclein expression may increase the risk of developing typical sporadic PD, whereas higher expression may cause severe forms of Parkinsonism similar to DLB. Pathologically, the burden of LB correlates with a PD or DLB clinical diagnosis, and it is still unclear whether PD and DLB are a continuum within the disease spectrum.

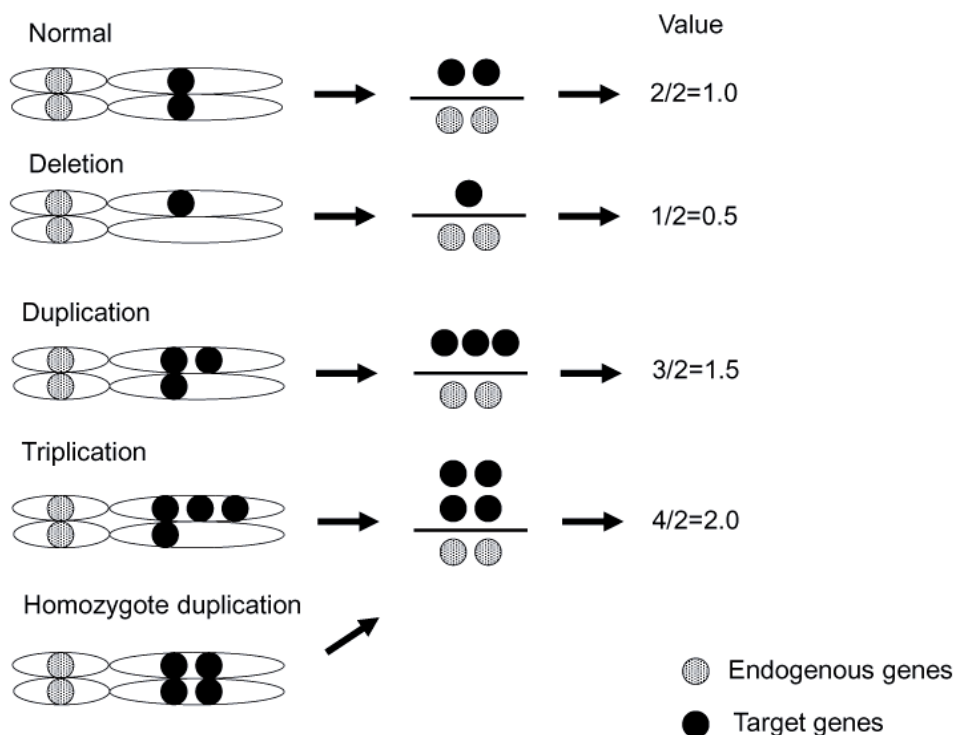


Fig. 1. A model for gene dosage of deletion, duplication and triplication

Singleton et al. first reported *SNCA* gene triplication within a large family with PD and dementia family (the Iowa kindred) (Singleton et al., 2003). The family members had been followed both clinically and pathologically by Mayo Clinic doctors for 100 years (Muentert et al., 1998, Gwinn-Hardy et al., 2000). The pattern of inheritance is autosomal dominant. The patients show a severe clinical course, early age at onset, complicating with dementia, or Parkinsonism. The pathological features are similar to DLB; (1) widespread LB pathology in the brain, (2) neuronal loss in the CA2/3, and (3) neuronal loss in the substantia nigra (Muentert et al., 1998). *SNCA* triplication was confirmed with quantitative PCR and FISH methodology. The size of the triplication region was over 2.0Mb. The expression values of messenger RNA and protein in peripheral blood and brain were twice the amount that is present in control subjects as predicted (Miller et al., 2004). Apart from the Iowan family, a

Swedish-American family was also reported as a *SNCA* triplication family and the patients had DLB-type prognosis and pathological findings. Their expression level of messenger RNA and protein of alpha-synuclein was also double the dosage of normal subjects in frontal cortex (Farrer et al., 2004).

These findings provided two novel insights regarding the underlying mechanisms of PD; (1) over-expression of alpha-synuclein may cause a more severe form of Parkinsonism such as DLB and (2) the gene dosage of alpha-synuclein may directly correlate with the clinical features of PD. Furthermore, these findings give us hints into the development of potential therapeutic avenues of treatment for PD by decreasing the expression of alpha-synuclein. The suppression of alpha-synuclein by siRNA approaches has proven successful in decreasing the levels of LB pathology in animal models (Lewis et al., 2008, McCormack et al., 2010).

After the detection of *SNCA* locus triplication, *SNCA* duplications were also reported in two families from France and Italy (Chartier-Harlin et al., 2004, Ibanez et al., 2004). Chartier-Harlin et al detected one duplication family among nine ADPD. Ibanez et al detected two families with *SNCA* duplication among 119 ADPD by 250K Affymetrix microarray and semi-quantitative PCR method (Ibanez et al., 2004). The symptoms of the patients with *SNCA* duplication are milder than that of *SNCA* triplication, younger age at onset and have a good efficacy for levodopa therapy, similar to sporadic PD.

Following these findings, we started screening for *SNCA* multiplications within 216 ADPD and 271 sporadic PD patients originating from Japan (Nishioka et al., 2006, Nishioka et al., 2009). We found six ADPD families and one sporadic case with *SNCA* duplication. Haplotype analysis showed these seven patients are derived from four common founders. Interestingly, the clinical manifestations of these patients were quite diverse such as sporadic PD, DLB-type and also many elderly asymptomatic carriers. The estimated penetrance ratio is about 30-40%. One patient presented with a severe clinical course with no efficacy for levodopa therapy. He progressed to Hoehn and Yahr stage V in a few years and at autopsy demonstrated features similar to DLB (Obi et al., 2008).

In addition, we also detected five asymptomatic carriers over 65 years old. We therefore focused our work on the reasons why the asymptomatic carriers at later ages do not present any clinical features of parkinsonism. We started to assess Brain MRI, PET study with [ $^{18}\text{F}$ ]-labeled 2-fluoro-2-deoxy-D-glucose (FDG) and [ $^{11}\text{C}$ ]-labeled 2 $\beta$ -carbomethoxy-3 $\beta$ -(4-fluorophenyl)-tropane (CFT), polysomnography, and Sniffin' sticks to explore the differences between the patients and the asymptomatic carriers with *SNCA* duplication. Against our expectation, the assessments for asymptomatic carriers did not show any abnormal results in our studies, with similar results as those obtained for normal individuals (Nishioka et al., 2009). It is an area of intense research why the asymptomatic carriers do not develop Parkinsonism at these late ages and its resolution will be a key in the puzzle regarding the late-onset of PD.

Apart from our cases, other teams have reported two families from Japan with *SNCA* duplication (Ikeuchi et al., 2008, Uchiyama et al., 2008). Interestingly, one family has both heterozygote and homozygote duplication (producing a pseudo-triplication) (Ikeuchi et al., 2008). The clinical features of the individual with the *SNCA* homozygote duplication showed severe Parkinsonism similar to that of a triplication carrier. These findings also confirm the gene dosage effect within the same family. Earlier studies had reported that the

Swedish family named the “Lister family complex” has both *SNCA* duplication and triplication patients within different branches of the pedigree suggesting a primary duplication event followed later by another resulting in the triplication (Fuchs et al., 2007). In this family also the patient with *SNCA* triplication presented with more severe symptoms than the patients with duplication. Recently one small pedigree with *SNCA* triplication was detected in Japan (Sekine et al., 2010).

The breakpoint of *SNCA* multiplication is different in each family. The largest multiplication about 4.9Mb is detected within a French family. The smallest one about 0.2 Mb is in a Japanese family (Nishioka et al., 2009). The size and gene make-up of each multiplication region does not seem to influence the clinical presentation of the carrier. The single common determining factor that appears between all patients with *SNCA* multiplication is the presence of the entire *SNCA* gene. To conclude, it is clear that *SNCA* multiplication alone is sufficient to result in the parkinsonian phenotype.

### 2.1 Sporadic PD and *SNCA* duplication

In 2007, Ahn et al reported two sporadic patients with *SNCA* duplication from a screen of 906 PD patients (Ahn et al., 2008). The age at onset was 65 and 50 years old for the two patients. Their clinical course was similar to typical sporadic PD without severe progression or cognitive decline. The estimated penetrance ratio was 33.3% among the Korean patients. Our studies have also detected one sporadic patient with *SNCA* duplication and this means that the low penetrance *SNCA* duplications may give the appearance of sporadic disease.

### 2.2 The frequency of *SNCA* multiplications in PD

The prevalence of *SNCA* multiplications is relatively low (table 1). Bruggemann et al reported one sporadic case among 403 PD cases from Germany ( $1/403=0.25\%$ ) (Bruggemann et al., 2008). Troiano et al reported one sporadic cases among 101 young onset PD cases from French ( $1/101=1\%$ ) (Troiano et al., 2008). Nuytemans et al reported one duplication patient with dementia among 219 sporadic PD cases from Belgium (Nuytemans et al., 2009). Sironi et al reported one duplication patient with dementia among 144 PD cases from Italy ( $1/144=0.7\%$ ) (Sironi et al., 2009). Furthermore some reports did not detect *SNCA* duplication; 0/50 and 0/290 (Hope et al., 2004, Xiromerisiou et al., 2007). To conclude, *SNCA* multiplication is not a common cause of sporadic or hereditary PD.

### 2.3 *SNCA* multiplications in multiple system atrophy

Multiple system atrophy (MSA) is characterized by specific clinical features such as Parkinsonism, autonomic dysfunction, poor response to levodopa therapy, and cerebellar ataxia (Wenning et al., 2004). Glial cytoplasmic inclusions (GCIs) are the pathologic hallmark of the disease. As alpha-synuclein is a major protein component of GCIs, MSA is categorized within the group of neurodegenerative disorders classified as the alpha-synucleinopathies. Interestingly common variation at the *SNCA* locus has been associated with MSA risk (Scholz et al., 2009, Ross et al., 2010). Two studies did not identify any *SNCA* multiplications in a combined total of 258 MSA patients (Lincoln et al., 2007, Ahn et al., 2008). Although the number of assessed samples may be small, these findings suggest that *SNCA* dosage is not a common cause of MSA. It is speculated that PD or DLB may be caused by lysosomal dysfunction, however, MSA may be caused by the oligodendrocytic changes in myelin basic protein (Wenning et al., 2008).



	Country	The number of pedigrees	The number of screening samples	Frequency (%)	Average age at onset
<b>SNCA duplication in AD cases</b>					
Nishioka et al. 2006 and 2009	Japan (Juntendo)	6	487	1.2	48.2
Ibanez et al. 2004 and 2009	France and Italy	4	286	1.4	43.6
Sironi et al. 2009	Italy	1	144	0.7	41
Ikeuchi et al. 2008	Japan (Niigata)	1			57
Uchiyama et al. 2008	Japan (Niigata)	1			60
Fuchs et al. 2007	Sweden (Lister complex)	1			71
Ahn et al. 2007	Korea	1	906	0.1	40
Chartier-Harlin et al. 2004	France	1	9	11	50.8
total average					51.5
<b>SNCA duplication in sporadic cases</b>					
Nishioka et al. 2009	Japan (Juntendo)	1	487	0.2	31
Nuytemans et al. 2009	Belgium	1	219	0.5	71
Brueggemann et al. 2008	Germany	1	403	0.3	36
Troiano et al. 2008	France	1	101	1	35
Ahn et al. 2007	Korea	2	906	0.2	57.5
total average					46.1
<b>SNCA triplication</b>					
Sekine et al. 2010	Japan (Juntendo)	1			37
Ibanez et al. 2009	France	1	286	0.3	42
Fuchs et al. 2007	Sweden (Lister complex)	1			
Farrer et al. 2004	Swedish-American	1			31
Singlton et al. 2003	Iowa	1			33.2
<b>SNCA homozygote duplication</b>					
Ikeuchi et al. 2008	Japan (Niigata)	1			28
total average					32.9

Table 1. The clinical manifestations and prevalence of SNCA duplication and triplication

## 2.4 The Synuclein family in Parkinson disease

The *SNCA* gene is located on chromosome 4q21-22 and is associated with susceptibility to PD and DLB. Alpha-synuclein has two paralogous genes, beta- (*SNCB*; MIM#602569) and gamma-synuclein (*SNCG*; MIM#602998) with which it shares a highly conserved N-terminal domain. *SNCB* is located on chromosome 5q35, and *SNCG* is located on chromosome 10q23 associated with breast and ovarian cancer (Ji et al., 1997, Goedert, 2001). All three synuclein genes are highly expressed in brain; thalamus, substantia nigra, caudate nucleus, and amygdala (Lavedan, 1998, Lavedan et al., 1998). A phylogenetic tree indicates that alpha- and beta- synucleins are related more closely to each other than to gamma-synuclein (Lavedan, 1998). Interestingly, two putative pathogenic mutations in *SNCB* are reported to cause DLB, however no significant co-segregation with disease could be shown and no other studies have identified these variants (Ohtake et al., 2004). A murine model with over-expressed gamma-synuclein is reported as a PD model with motor deficits (Ninkina et al., 2009). Our recent studies on common variation in the synuclein family of genes also suggested association for variants in both *SNCA* and *SNCG* with diffuse LB disease (Nishioka et al., 2010). Given these findings, it is postulated that there is a connection between not only *SNCA*, but also *SNCB* and *SNCG* and susceptibility to PD, however multiplications of the *SNCB* and *SNCG* loci have not yet been observed.

## 3. Conclusion and future work

Research focused on copy number variation has made remarkable progress in recent years. Genome-wide studies for copy number variants (CNV) indicate 1447 copy number variable regions (CNVRs) (Redon et al., 2006). Presumably, many of these CNV polymorphisms result in differential expression levels of proteins and dictate the phenotypic presentation at the individual level. Interestingly in Alzheimer disease multiplications of the *APP* gene have also been identified in families with autosomal dominantly inherited forms of the disease (Cabrejo et al., 2006, Rovelet-Lecrux et al., 2006). Robust and comprehensive studies are now warranted for CNV across the genome and may not only help develop new treatments for PD but perhaps several other neurodegenerative diseases.

## 4. References

- Ahn TB, Kim SY, Kim JY, Park SS, Lee DS, Min HJ, Kim YK, Kim SE, Kim JM, Kim HJ, Cho J, Jeon BS (2008) alpha-Synuclein gene duplication is present in sporadic Parkinson disease. *Neurology* 70:43-49.
- Braak H, Rub U, Gai WP, Del Tredici K (2003) Idiopathic Parkinson's disease: possible routes by which vulnerable neuronal types may be subject to neuroinvasion by an unknown pathogen. *J Neural Transm* 110:517-536.
- Brueggemann N, Odin P, Gruenewald A, Tadic V, Hagenah J, Seidel G, Lohmann K, Klein C, Djarmati A (2008) Re: Alpha-synuclein gene duplication is present in sporadic Parkinson disease. *Neurology* 71:1294; author reply 1294.
- Cabrejo L, Guyant-Marchal L, Laquerriere A, Vercelletto M, De la Fourniere F, Thomas-Anterion C, Verny C, Letournel F, Pasquier F, Vital A, Checler F, Frebourg T, Campion D, Hannequin D (2006) Phenotype associated with APP duplication in five families. *Brain* 129:2966-2976.

- Chartier-Harlin MC, Kachergus J, Roumier C, Mouroux V, Douay X, Lincoln S, Levecque C, Larvor L, Andrieux J, Hulihan M, Waucquier N, Defebvre L, Amouyel P, Farrer M, Destee A (2004) Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet* 364:1167-1169.
- Farrer M, Kachergus J, Forno L, Lincoln S, Wang DS, Hulihan M, Maraganore D, Gwinn-Hardy K, Wszolek Z, Dickson D, Langston JW (2004) Comparison of kindreds with parkinsonism and alpha-synuclein genomic multiplications. *Ann Neurol* 55:174-179.
- Farrer MJ (2006) Genetics of Parkinson disease: paradigm shifts and future prospects. *Nat Rev Genet* 7:306-318.
- Fuchs J, Nilsson C, Kachergus J, Munz M, Larsson EM, Schule B, Langston JW, Middleton FA, Ross OA, Hulihan M, Gasser T, Farrer MJ (2007) Phenotypic variation in a large Swedish pedigree due to SNCA duplication and triplication. *Neurology* 68:916-922.
- Goedert M (2001) Alpha-synuclein and neurodegenerative diseases. *Nat Rev Neurosci* 2:492-501.
- Gwinn-Hardy K, Mehta ND, Farrer M, Maraganore D, Muentner M, Yen SH, Hardy J, Dickson DW (2000) Distinctive neuropathology revealed by alpha-synuclein antibodies in hereditary parkinsonism and dementia linked to chromosome 4p. *Acta Neuropathol (Berl)* 99:663-672.
- Hope AD, Myhre R, Kachergus J, Lincoln S, Bisceglia G, Hulihan M, Farrer MJ (2004) Alpha-synuclein missense and multiplication mutations in autosomal dominant Parkinson's disease. *Neuroscience letters* 367:97-100.
- Ibanez P, Bonnet AM, Debarges B, Lohmann E, Tison F, Pollak P, Agid Y, Durr A, Brice A (2004) Causal relation between alpha-synuclein gene duplication and familial Parkinson's disease. *Lancet* 364:1169-1171.
- Ikeuchi T, Kakita A, Shiga A, Kasuga K, Kaneko H, Tan CF, Idezuka J, Wakabayashi K, Onodera O, Iwatsubo T, Nishizawa M, Takahashi H, Ishikawa A (2008) Patients homozygous and heterozygous for SNCA duplication in a family with parkinsonism and dementia. *Archives of neurology* 65:514-519.
- Ji H, Liu YE, Jia T, Wang M, Liu J, Xiao G, Joseph BK, Rosen C, Shi YE (1997) Identification of a breast cancer-specific gene, BCSG1, by direct differential cDNA sequencing. *Cancer Res* 57:759-764.
- Kruger R, Kuhn W, Muller T, Woitalla D, Graeber M, Kosel S, Przuntek H, Epplen JT, Schols L, Riess O (1998) Ala30Pro mutation in the gene encoding alpha-synuclein in Parkinson's disease. *Nat Genet* 18:106-108.
- Lavedan C (1998) The synuclein family. *Genome Res* 8:871-880.
- Lavedan C, Leroy E, Torres R, Dehejia A, Dutra A, Buchholtz S, Nussbaum RL, Polymeropoulos MH (1998) Genomic organization and expression of the human beta-synuclein gene (SNCB). *Genomics* 54:173-175.
- Lewis J, Melrose H, Bumcrot D, Hope A, Zehr C, Lincoln S, Braithwaite A, He Z, Ogholikhan S, Hinkle K, Kent C, Toudjarska I, Charisse K, Braich R, Pandey RK, Heckman M, Maraganore DM, Crook J, Farrer MJ (2008) In vivo silencing of alpha-synuclein using naked siRNA. *Mol Neurodegener* 3:19.
- Lincoln SJ, Ross OA, Milkovic NM, Dickson DW, Rajput A, Robinson CA, Papapetropoulos S, Mash DC, Farrer MJ (2007) Quantitative PCR-based screening of alpha-synuclein

- multiplication in multiple system atrophy. *Parkinsonism & related disorders* 13:340-342.
- McCormack AL, Mak SK, Henderson JM, Bumcrot D, Farrer MJ, Di Monte DA (2010) Alpha-synuclein suppression by targeted small interfering RNA in the primate substantia nigra. *PLoS One* 5:e12122.
- McKeith IG, Dickson DW, Lowe J, Emre M, O'Brien JT, Feldman H, Cummings J, Duda JE, Lippa C, Perry EK, Aarsland D, Arai H, Ballard CG, Boeve B, Burn DJ, Costa D, Del Ser T, Dubois B, Galasko D, Gauthier S, Goetz CG, Gomez-Tortosa E, Halliday G, Hansen LA, Hardy J, Iwatsubo T, Kalaria RN, Kaufer D, Kenny RA, Korczyn A, Kosaka K, Lee VM, Lees A, Litvan I, Londos E, Lopez OL, Minoshima S, Mizuno Y, Molina JA, Mukaetova-Ladinska EB, Pasquier F, Perry RH, Schulz JB, Trojanowski JQ, Yamada M (2005) Diagnosis and management of dementia with Lewy bodies: third report of the DLB Consortium. *Neurology* 65:1863-1872.
- Miller DW, Hague SM, Clarimon J, Baptista M, Gwinn-Hardy K, Cookson MR, Singleton AB (2004) Alpha-synuclein in blood and brain from familial Parkinson disease with SNCA locus triplication. *Neurology* 62:1835-1838.
- Muenter MD, Forno LS, Hornykiewicz O, Kish SJ, Maraganore DM, Caselli RJ, Okazaki H, Howard FM, Jr., Snow BJ, Calne DB (1998) Hereditary form of parkinsonism--dementia. *Ann Neurol* 43:768-781.
- Ninkina N, Peters O, Millership S, Salem H, van der Putten H, Buchman VL (2009) Gamma-synucleinopathy: neurodegeneration associated with overexpression of the mouse protein. *Human molecular genetics* 18:1779-1794.
- Nishioka K, Hayashi S, Farrer MJ, Singleton AB, Yoshino H, Imai H, Kitami T, Sato K, Kuroda R, Tomiyama H, Mizoguchi K, Murata M, Toda T, Imoto I, Inazawa J, Mizuno Y, Hattori N (2006) Clinical heterogeneity of alpha-synuclein gene duplication in Parkinson's disease. *Ann Neurol* 59:298-309.
- Nishioka K, Ross OA, Ishii K, Kachergus JM, Ishiwata K, Kitagawa M, Kono S, Obi T, Mizoguchi K, Inoue Y, Imai H, Takanashi M, Mizuno Y, Farrer MJ, Hattori N (2009) Expanding the clinical phenotype of SNCA duplication carriers. *Mov Disord* 24:1811-1819.
- Nishioka K, Wider C, Vilarino-Guell C, Soto-Ortolaza AI, Lincoln SJ, Kachergus JM, Jasinska-Myga B, Ross OA, Rajput A, Robinson CA, Ferman TJ, Wszolek ZK, Dickson DW, Farrer MJ (2010) Association of alpha-, beta-, and gamma-Synuclein with diffuse lewy body disease. *Archives of neurology* 67:970-975.
- Nuytemans K, Meeus B, Crosiers D, Brouwers N, Goossens D, Engelborghs S, Pals P, Pickut B, Van den Broeck M, Corsmit E, Cras P, De Deyn PP, Del-Favero J, Van Broeckhoven C, Theuns J (2009) Relative contribution of simple mutations vs. copy number variations in five Parkinson disease genes in the Belgian population. *Human mutation* 30:1054-1061.
- Obi T, Nishioka K, Ross OA, Terada T, Yamazaki K, Sugiura A, Takanashi M, Mizoguchi K, Mori H, Mizuno Y, Hattori N (2008) Clinicopathologic study of a SNCA gene duplication patient with Parkinson disease and dementia. *Neurology* 70:238-241.
- Ohtake H, Limprasert P, Fan Y, Onodera O, Kakita A, Takahashi H, Bonner LT, Tsuang DW, Murray IV, Lee VM, Trojanowski JQ, Ishikawa A, Idezuka J, Murata M, Toda T, Bird TD, Leverenz JB, Tsuji S, La Spada AR (2004) Beta-synuclein gene alterations in dementia with Lewy bodies. *Neurology* 63:805-811.

- Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodward C, Yang F, Zhang J, Zerjal T, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME (2006) Global variation in copy number in the human genome. *Nature* 444:444-454.
- Ross OA, Vilarino-Guell C, Wszolek ZK, Farrer MJ, Dickson DW (2010) Reply to: SNCA variants are associated with increased risk of multiple system atrophy. *Ann Neurol* 67:414-415.
- Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerriere A, Vital A, Dumanchin C, Feuillet S, Brice A, Vercelletto M, Dubas F, Frebourg T, Campion D (2006) APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat Genet* 38:24-26.
- Scholz SW, Houlden H, Schulte C, Sharma M, Li A, Berg D, Melchers A, Paudel R, Gibbs JR, Simon-Sanchez J, Paisan-Ruiz C, Bras J, Ding J, Chen H, Traynor BJ, Arepalli S, Zonozi RR, Revesz T, Holton J, Wood N, Lees A, Oertel W, Wullner U, Goldwurm S, Pellecchia MT, Illig T, Riess O, Fernandez HH, Rodriguez RL, Okun MS, Poewe W, Wenning GK, Hardy JA, Singleton AB, Del Sorbo F, Schneider S, Bhatia KP, Gasser T (2009) SNCA variants are associated with increased risk for multiple system atrophy. *Ann Neurol* 65:610-614.
- Sekine T, Kagaya H, Funayama M, Li Y, Yoshino H, Tomiyama H, Hattori N (Clinical course of the first Asian family with Parkinsonism related to SNCA triplication. *Mov Disord* 25:2871-2875.2010).
- Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, Kachergus J, Hulihan M, Peuralinna T, Dutra A, Nussbaum R, Lincoln S, Crawley A, Hanson M, Maraganore D, Adler C, Cookson MR, Muentner M, Baptista M, Miller D, Blancato J, Hardy J, Gwinn-Hardy K (2003) alpha-Synuclein locus triplication causes Parkinson's disease. *Science* 302:841.
- Sironi F, Trotta L, Antonini A, Zini M, Ciccone R, Della Mina E, Meucci N, Sacilotto G, Primignani P, Brambilla T, Coviello DA, Pezzoli G, Goldwurm S (2009) alpha-Synuclein multiplication analysis in Italian familial Parkinson disease. *Parkinsonism & related disorders*.
- Troiano AR, Cazeneuve C, Le Ber I, Bonnet AM, Lesage S, Brice A (2008) Re: Alpha-synuclein gene duplication is present in sporadic Parkinson disease. *Neurology* 71:1295; author reply 1295.
- Uchiyama T, Ikeuchi T, Ouchi Y, Sakamoto M, Kasuga K, Shiga A, Suzuki M, Ito M, Atsumi T, Shimizu T, Ohashi T (2008) Prominent psychiatric symptoms and glucose hypometabolism in a family with a SNCA duplication. *Neurology* 71:1289-1291.
- Wenning GK, Colosimo C, Geser F, Poewe W (2004) Multiple system atrophy. *Lancet Neurol* 3:93-103.
- Wenning GK, Stefanova N, Jellinger KA, Poewe W, Schlossmacher MG (2008) Multiple system atrophy: a primary oligodendrogliopathy. *Ann Neurol* 64:239-246.
- Xiromerisiou G, Hadjigeorgiou GM, Gourbali V, Johnson J, Papakonstantinou I, Papadimitriou A, Singleton AB (2007) Screening for SNCA and LRRK2 mutations

in Greek sporadic and autosomal dominant Parkinson's disease: identification of two novel LRRK2 variants. *Eur J Neurol* 14:7-11.

Zarranz JJ, Alegre J, Gomez-Esteban JC, Lezcano E, Ros R, Ampuero I, Vidal L, Hoenicka J, Rodriguez O, Atares B, Llorens V, Gomez Tortosa E, del Ser T, Munoz DG, de Yebenes JG (2004) The new mutation, E46K, of alpha-synuclein causes Parkinson and Lewy body dementia. *Ann Neurol* 55:164-173.

# Bucentaur (*Bcnt*)<sup>1</sup> Gene Family: Gene Duplication and Retrotransposon Insertion

Shintaro Iwashita and Naoki Osada  
*Iwaki Meisei University / National Institute of Genetics*  
Japan

## 1. Introduction

Members of multiple gene families in higher organisms allow for more refined cellular signaling networks and structural organization toward more stable physiological homeostasis. Gene duplication is one the most powerful ways of providing an opportunity to create a novel gene(s) because a novel function might be acquired without the loss of the original gene function (Ohno, 1970). Gene duplication can result from unequal crossing over by recombination, retroposition of cDNA, or whole-genome duplication. Furthermore, a replication-based mechanism of change in gene copy number has been proposed recently (Hastings et al., 2009). Gene duplication generated by retroposition is frequently accompanied by deleterious effects because the insertion of cDNA into the genome is nearly random or unlinks the original gene location resulting in an alteration of the original vital functions of the target genes. Thus retroelements such as transposable elements and endogenous retroviruses have been thought of as “selfish”. On the other hand, gene duplication caused by unequal crossing over generally results in tandem alignment, which less frequently disrupts the functions of other genes. Recent genome-wide studies have demonstrated that retroelements can definitely contribute to the creation of individual novel genes and the modulation of gene expression, which allows for the dynamic diversity of biological systems, such as placental evolution (Rawn & Cross, 2008). It is now recognized that tandem duplication and retroposition are among the key factors that initiate the creation of novel gene family members (Brosius, 2005; Sorek, 2007; Kaessmann, 2010). By these mechanisms, species-specific gene duplication can lead to species-specific gene functions, which might contribute to species-specific phenotypes (Zhang, 2003). For example, many genes derived from retroelements are expressed in mammalian placentas, and species-specific gene duplication has occurred multiple times during placental evolution (Rawn & Cross, 2008). If a combination of tandem gene duplication and retroposition of cDNA occurs, there is a good possibility for the creation of a novel gene(s)

---

<sup>1</sup>Although the vertebrate *Bcnt* (Bucentaur) gene is officially called *Cfdp1* (craniofacial developmental protein 1), its biological function remains unclear. So far, solid evidence that the gene is involved in craniofacial development has not been provided except for its unique expression during mouse tooth development (Diekwisch et al., 1999). The authors are concerned that a “wrong” naming may have caused confusion concerning the function of the *Bcnt*/*Cfdp1* gene. Thus we use the names *Bcnt*/*Cfdp1*, *p97Bcnt*/*Cfdp2* and *p97Bcnt-2* in this article.

because a novel function could be acquired with the guarantee that the original gene functions will be retained. This type of evolutionary process has been described, e.g. in the *Jingwei* gene of *Drosophila*, where segmental duplication of a certain gene followed by retroposition of alcohol dehydrogenase (*Adh*) cDNA into one of the copied genes created a new *Adh* with an altered substrate specificity (Zhang et al., 2004). Furthermore, since the insertion of retrotransposons can speed up the natural mutation process tremendously (Makalowski, 2003), the combined process of tandem duplication followed by retrotransposon insertion has a greater potential to generate a novel gene(s) (Fig. 1). Indeed we have identified an example of this type, the p97Bcnt protein (Nobukuni et al., 1997). The *p97Bcnt/Cfdp2* gene was created in a common ancestor of ruminants by a partial duplication of the ancestral *Bcnt/Cfdp1* gene followed by insertion of an order-specific retrotransposon, Bov B-LINE (Iwashita et al., 2003, 2006). As a result, the paralog recruited an apurinic/apyrimidinic (AP)-endonuclease domain in the middle of the protein. In this article, we summarize the gene organization and protein structures of three Bcnt family members, and describe their biochemical characteristics. We also argue that the process of tandem duplication followed by retroelement insertion generates a high potential for creating novel genes for expanding signaling networks.

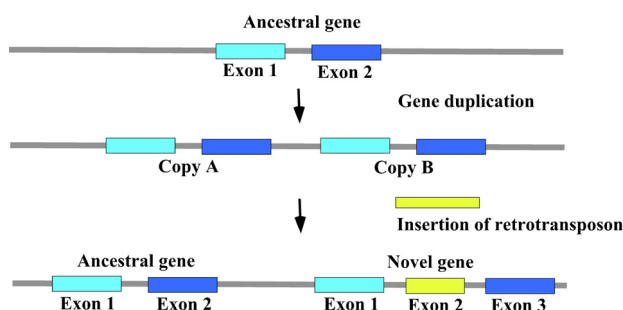


Fig. 1. Mechanism of novel gene creation by a combination of gene duplication and retrotransposition

If segmental duplication followed by retrotransposon insertion occurs, it provides a good opportunity to generate a paralogous gene, because a novel function could be acquired under the guarantee that the original gene function will remain. The schematic is a modification of the original (Makalowski, 2003)

## 2. Establishment of three *Bcnt* members

### 2.1 A bovine specific retrotransposon, Bov B-LINE

Autonomous non-long-terminal repeat (non-LTR) retrotransposons, also termed Long-Interspersed Nuclear Elements (LINEs), have been identified in almost all eukaryotic organisms. Based on their structures and type of endonuclease, non-LTR retrotransposons are classified into two subtypes. The major subtype encodes an endonuclease with homology to AP-endonuclease (APE), thus termed APE-type non-LTR retrotransposons. These APE-type elements are now divided into four groups and eleven clades (Zingler et al., 2005). The RTE clade is one of the most widespread and shortest APE-type non-LTR retrotransposons, which are truncated forms of L1 (human LINE 1) lacking both the 5' and 3' regions (Fig. 2).



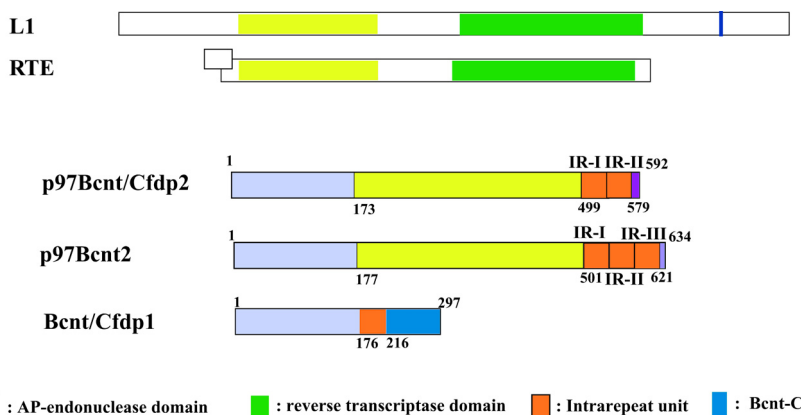


Fig. 2. The structural relationships among the open reading frames of retrotransposable L1 and RTE elements and three Bcnt-related proteins

The L1 and RTE elements have apurinic/apyrimidic (AP)-endonuclease domains (yellow boxes) and reverse transcriptase domains (green boxes). The L1 element has another restriction enzyme-like endonuclease domain in the C-terminal region (dark blue bar). The square to the left of RTE indicates the ambiguity of the 5' region. The assignment of the domains in L1 and RTE is according to Malik & Eickbush, 1998. The numbers above the rectangles of the three Bcnt-related proteins, Bcnt/Cfdp1, p97Bcnt/Cfdp2 and p97Bcnt2, indicate amino acid residue numbers. The latter two contain a region derived from the AP-endonuclease domain of RTE (termed the RTE domain) in the middle of their molecules. As described below, the three proteins have common acidic N-terminal regions (grey boxes) and intramolecular repeat (IR) units consisting of 40 amino acids each (orange boxes). The blue box at the C-terminus of ancestral Bcnt/Cfdp1 indicates a conserved 82-amino acids region (Bcnt-C)

Retrotransposons spread through vertical transmission, but occasionally through horizontal transmission (Gentles et al., 2007). Bov-B LINEs are order-specific RTEs that are found specifically in ruminants, where they were initially reported as a bovine *Alu*-like dimer-driven family; they potentially encode both an AP-endonuclease domain and a reverse transcriptase domain accompanied by a short interspersed repetitive element (SINE) cassette (Szemraj et al., 1995). It has been suggested that BovB-LINEs were transferred horizontally from squamata into an ancestral ruminant and expanded in all ruminants (Zupunski et al., 2001). *p97Bcnt/Cfdp2* recruited the AP-endonuclease domain of Bov-B LINE during the creation process in an ancient ruminant.

## 2.2 Discovery of a novel protein, p97Bcnt/Cfdp2

The p97Bcnt/Cfdp2 protein was discovered in bovine brain during screening for hybridoma producing monoclonal antibodies (mAbs). In the course of a study on Ras GTPase-activating proteins (GAPs, RAS p21 protein activators, Rasa), we had attempted to generate mAbs to distinguish each GAP from among their family members (Kobayashi et al., 1993; Iwashita & Song, 2008). We used a glutathione-S-transferase (GST) fusion protein of rat Rasa2 (GAP<sup>1m</sup>) as an immunoantigen and screened for hybridomas by western blotting using bovine brain extract. We isolated five independent clones, all of which showed a single broad band with an apparent molecular mass of 97 kDa, exactly the expected size of rat Rasa2. At one time

we thought we had obtained appropriate antibodies, but the target protein was entirely different from Rasa as described below. Although we screened a bovine brain cDNA expression library by western blotting with the obtained mAbs, we could not clone the target molecule. Instead, a 97 kDa protein was isolated from bovine brain extract by affinity chromatography with the antibodies, and the amino acid sequences of its protease-digested fragments were determined. We used redundant primers designed based on the determined peptide sequences as DNA probes, and cloned the target molecule by both "rapid amplification of cDNA ends" (RACE) and screening of a bovine brain cDNA library (Nobukuni et al., 1997). The obtained clone, which had an open reading frame of 592 amino acids, was named Bcnt after bucentaur, a Greek mythical creature that is half man and half ox, implying a strange protein from bovine brain. The identified protein, named p97Bcnt, Bcnt with a molecular mass of 97kDa, consists of an acidic N-terminal region, a retrotransposon-derived 325-amino acid region (termed the RTE domain), and two 40-amino acid intrarepeat (IR) units. The RTE domain is 72% identical to an order-specific retrotransposon, Bov-B LINE (GenBank accession number AF332697). The relationship between p97Bcnt and its estimated epitope in the mAbs, which enabled us to identify the protein, is summarized in Fig. 3. It provides a reasonable explanation as to why the unique protein was isolated by mAbs generated by a GST-fusion protein of rat Rasa2 as an immunoantigen. The estimated epitope of five independent mAbs maps on a single site in the N-terminal region of p97Bcnt/Cfdp2, and the antibodies recognize neither human BCNT/CFDP1 (Nobukuni et al., 1997) nor bovine Bcnt/Cfdp1 (Iwashita et al., 2003). The junction region of the fusion protein between GST and truncated Rasa2 codes a unique amino acid sequence generated by the extra nucleotides of the multiple cloning sites and a nucleotide linker for plasmid construction. Since Rasa is a highly conserved protein in mammals, the junction region might present strong antigenicity. Generally, it is hard to clone a target molecule by direct DNA screening when interspersed repetitive sequences are involved. Therefore, we first isolated a 97kDa protein that was recognized by the accidentally generated mAbs, determined its amino acid sequence, and then screened a cDNA library with the designed oligonucleotide probes. This led to the identification of a unique protein, p97Bcnt/Cfdp2.

### 2.3 Identification of three Bcnt-related proteins

Immediately after the identification of p97Bcnt/Cfdp2, we isolated its human and mouse counterparts, and examined their differences from p97Bcnt/Cfdp2 at both the cDNA and genome levels (Nobukuni et al., 1997; Takahashi et al., 1998). The counterparts, called (ancestral) Bcnt/Cfdp1, have homologous acidic N-terminal regions and one IR unit of 40-amino acids, but lack the RTE domain. Instead they contain a highly conserved 82-amino acid region at the C-terminus that is not present in p97Bcnt/Cfdp2 (Fig. 2) as will be described below. Subsequently, we found that ruminants have both ancestral Bcnt/Cfdp1 and p97Bcnt/Cfdp2, while other animals have only Bcnt/Cfdp1. The pairwise sequence alignment of bovine and human genome DNA revealed that the region encompassing the gene was duplicated in two rounds in bovines (Iwashita et al., 2003). Although automated computational annotation predicted another homolog of p97Bcnt (LOC514131) in the bovine genome, its 5' UTR was different from the full-length cDNA that we isolated. Then we identified another paralog, termed p97Bcnt-2, in the adjacent region (Iwashita et al., 2009). The gene product, p97Bcnt2, is highly homologous to p97Bcnt/Cfdp2, comprising an acidic N-terminal region, a 324-amino acid RTE domain, and three IR units instead of the two in p97Bcnt/Cfdp2 in the C-terminal region (Fig. 2).

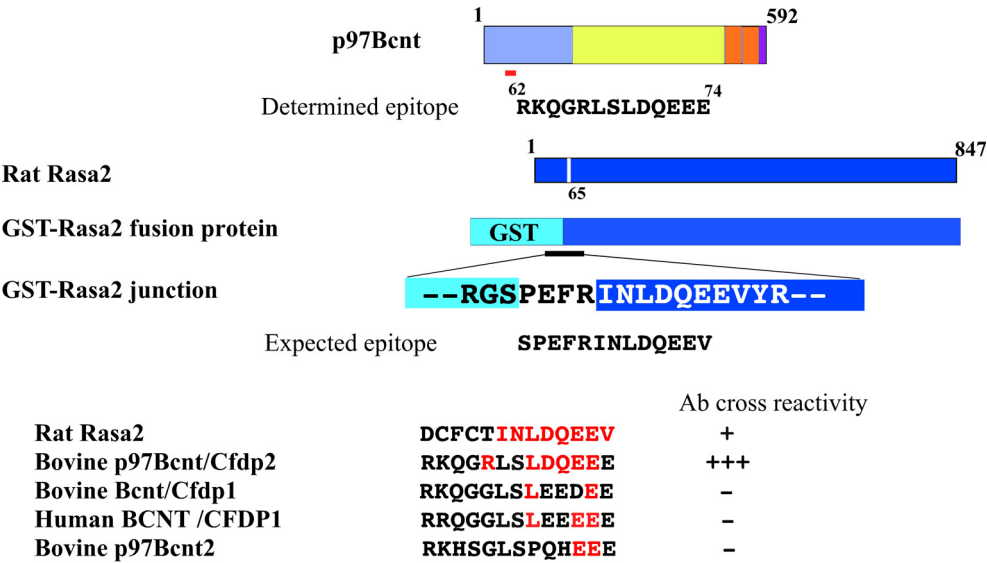


Fig. 3. Epitopes of the monoclonal antibodies that enabled the identification of the p97Bcnt/Cfdp2 protein

A plasmid of a fusion protein of truncated rat Rasa2 (from Ile65 to Ser847) and glutathione S-transferase was constructed using a linker (by Dr. S. Hattori), expressed in *Escherichia coli*, and its protein was purified by glutathione-affinity column chromatography. mAbs against the fusion protein were isolated according to a conventional method. Epitope mapping was carried out using the full-length cDNA of the targeted molecule, hereafter *p97Bcnt*. Fragments of ~300 base pairs in size were expressed in a protein expression vector and screened with the obtained mAbs. Seven positive clones were isolated from among ~9 × 10<sup>3</sup> bacterial colonies, and the sequence common to all clones was determined as the possible epitope for anti-p97Bcnt antibodies (13-amino acids, RKQGRSLDQEEE, represented by the red bar in the upper part) (Nobukuni et al., 1997). Amino acid sequences corresponding to the epitope region of rat Rasa2, bovine p97Bcnt/Cfdp2, human BCNT/CFDP1, bovine Bcnt/Cfdp1, and bovine p97Bcnt2 are aligned and amino acid residues identical to those in the expected epitope are indicated in red

3. Tandem alignment of three Bcnt gene family members

The draft bovine genome sequence was published in 2009 (The Bovine Genome Sequencing and Analysis Consortium, 2009). The initial analysis estimated that the bovine genome contains about 22,000 genes, with a core set of 14,345 orthologs shared among seven mammalian species. It has been shown that 3.1% of the bovine genome consists of recently duplicated sequences (judged by sequences ≥ 1 kb in length and ≥ 90% identity), and more than three-quarters (75-90%) of segmental duplications are organized into local tandem duplication clusters (Liu et al., 2009). It is noteworthy that cattle-specific evolutionary breakpoint regions in the chromosomes have a higher density of tandem duplications and enrichment of repetitive elements. Furthermore, it has been pointed out that bovine tandem gene duplication is significantly related to species-specific biological functions such as immunity, digestion, lactation, and reproduction (Liu et al., 2009).

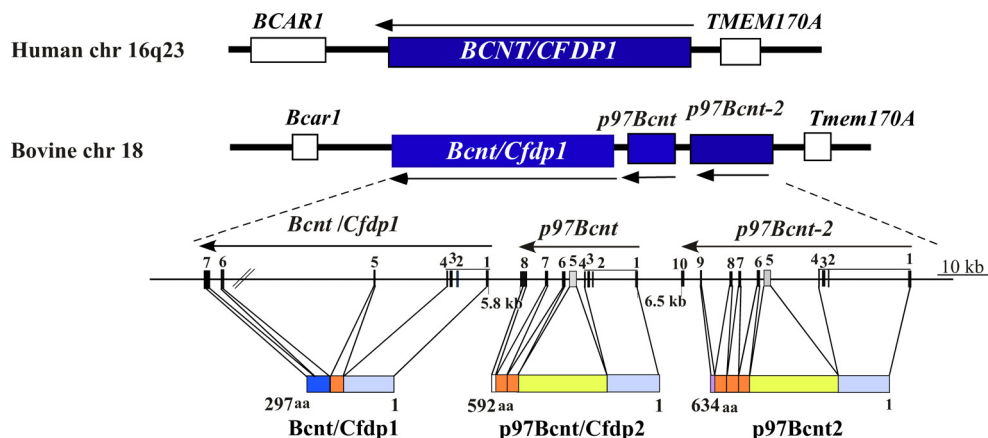


Fig. 4. Bovine *Bcnt/Cfdp1* locus and its corresponding region in the human genome

The organization of bovine *Bcnt/Cfdp1*-*p97Bcnt/Cfdp2*-*p97Bcnt-2* is shown schematically. *Bcar1*, Breast cancer anti-estrogen resistance 1 gene and *Tmem170A*, Transmembrane protein 170A gene, are located proximal and distal to the *Bcnt*-gene cluster or *BCNT* in both bovine chromosome 18 (middle part) and human chromosome 16q23 (upper part), respectively. The *Bcnt/Cfdp1*, *p97Bcnt/Cfdp2*, and *p97Bcnt-2* genes comprise 7, 8, and 10 exons, respectively (lower part); each exon is indicated by a vertical bar and is numbered

The three *Bcnt*-related genes are tandemly aligned on bovine chromosome 18 over a range of more than 177 kb, a syntenic region of human chromosome 16q23 (Fig. 4) and mouse chromosome 8. This gene cluster exists between the proximal breast cancer anti-estrogen resistance 1 gene (*Bcar1*) and the distal transmembrane protein 170A gene (*Tmem170A*) in bovines, as is the case of *BCNT/CFDP1* in humans and *Bcnt/Cfdp1* in mice. Therefore, the cluster region was generated from an order-specific segmental duplication. It has been suggested that Bov-B LINES emerged by horizontal transfer from squamata to ancient ruminants (Zupunski et al., 2001), and expanded just after the divergence of ruminantia and Camelidae (Jobse et al., 1995). Bov-B LINES have further expanded in different lineages during the diversification of ruminant species after splitting from Tragulina, which was confirmed by hybridization with DNA fragments of the RTE domain of *Lesser Malay chevrotain* (Iwashita et al., 2006). This is consistent with the expansion of bovine SINEs (Jobse et al., 1995). *Tragulus javanicus*, the living fossil of the basal ruminant stock, shares a similar *Bcnt/Cfdp1* and *p97Bcnt/Cfdp2* gene organization to bovines (Iwashita et al., 2006). Thus the partial gene duplication of the ancestral *Bcnt/Cfdp1* followed by the Bov-B LINE insertion occurred sometime after the Ruminantia-Suina-Tylopoda split and before the Pecora-Tragulina divergence, ~50 million years ago. A phylogenetic tree has been constructed based on the N-terminal regions (~175 amino acids) encoded by the first four exons and shared among three *Bcnt*-related members. The tree topology suggests *p97Bcnt-2* was created from duplication of ancestor *p97Bcnt/Cfdp2* in an ancient ruminant prior to the Pecora-Tragulina divergence. Furthermore, using the 120-bp sequence corresponding to 40 amino acid residues in IR, duplication of the IR unit in *p97Bcnt/Cfdp2* is estimated to have occurred prior to the creation of *p97Bcnt-2*, which has three IR units (Fig. 2). The two units in *p97Bcnt-2* (IR-

II and IR-III) diverged from IR-II in *p97Bcnt/Cfdp2*. We propose a parsimonious scenario for the creation of the three *Bcnt*-related genes in a process comprising 5 steps as shown in Fig. 5 (Iwashita et al., 2009). Self-BLAST search of the 120-kb region from *Tmem 170A* to exon 5 of *Bcnt/Cfdp1* confirms the two-round duplication of this gene cluster. Furthermore, homologous fragments of *Tmem170A* 3' UTR, which is located 6.8-kb distal to *p97Bcnt-2*, distribute at the 3'-region of both *Bcnt/Cfdp1* and *p97Bcnt/Cfdp2*. These data support the above scenario that resulted in the creation of the two paralogs, *p97Bcnt/Cfdp2* and *p97Bcnt-2*. Furthermore, both the processed pseudogene of *Bcnt/Cfdp1* and a 900-bp fragment encompassing the IR-II exon of *p9Bcnt-2* map on bovine chromosome 26 (Iwashita et al., 2009). It is interesting to examine the relationship between the retrotransposon-mediated creation of novel genes and the occurrence of processed pseudogenes.

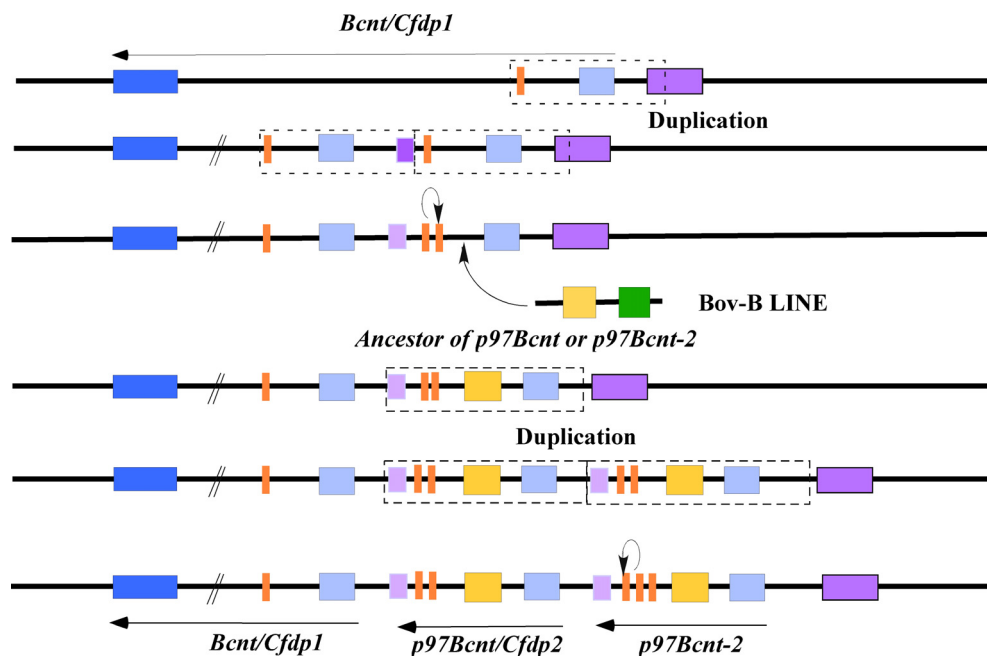


Fig. 5. A scenario for the creation of the two paralogous genes, *p97Bcnt/Cfdp2* and *p97Bcnt-2*

A parsimonious scenario for the creation of the three *Bcnt*-related family genes includes 5 steps as follows: (1) partial gene duplication of the ancestral *Bcnt/Cfdp1*, leaving the Bcnt-C region by segmental duplication; (2) insertion of a Bov-B LINE in intron 5 of one of the duplicated copies, recruitment of the AP-endonuclease domain of the retrotransposon, and generation of the ancestor of *p97Bcnt/Cfdp2* or *p97Bcnt-2*; (3) segmental duplication of the IR unit of ancestor *p97Bcnt*; (4) further gene duplication of the ancestor *p97Bcnt* to generate the nascent *p97Bcnt-2*; and, finally, (5) segmental duplication of the IR unit of the nascent *p97Bcnt-2* to create *p97Bcnt-2*. Nucleotide regions corresponding to the acidic N-terminal regions, IR units, Bcnt-C, and *Tmem170A*, are symbolically indicated by boxes colored grey, orange, dark blue, and purple, respectively. Bov-B LINE has an AP-endonuclease domain (in yellow) and reverse transcriptase domain (in green)

## 4. Characteristics of three *Bcnt*-related gene products

### 4.1 Ancestral *Bcnt* protein with highly conserved C-terminal region (*Bcnt*-C)

It has been proposed that duplicated genes yield genetic redundancy, which should result in either the acquisition of a gene with a novel function or the degeneration of one of the duplicated genes. Two paralogs, *p97Bcnt/Cfdp2* and *p97Bcnt-2* genes, were created via a process of tandem duplication followed by retrotransposon insertion. We expect that the three *Bcnt*-related proteins may play a role in more refined cellular signaling in ruminants. The vertebrate *Bcnt/Cfdp1* protein includes a highly conserved 82-amino acid region at the C-terminus, termed *Bcnt*-C, which is not present in either *p97Bcnt/Cfdp2* or *p97Bcnt2* (Iwashita et al., 2003; 2009) (Fig. 2). *Bcnt*-C, known as the BCNT superfamily, is found in most eukaryotes, including yeast, and is classified into Pfam 07572 in the Pfam database (<http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=pfam07572>). Although the functions of the BCNT family members remain mostly unclear, a vertebrate *Bcnt/Cfdp1* was recently identified as a centomere protein, CENP-29, in DT40 cells, a chicken B cell line transformed by avian leukosis virus (Ohta et al., 2010). Furthermore, a yeast ortholog, Swc5/YBR231C/AOR1, is a component of the chromatin-remodeling complex SWR1 in *Saccharomyces cerevisiae* (budding yeast) (Wu et al., 2009). The SWR1 complex mediates the ATP-dependent exchange of histone H2A for the H2A variant HZT1, and the Swc5 null mutant shows phenotypes of decreased resistance to macromolecule synthesis inhibitors such as hydroxyurea and cycloheximide, and increased heat sensitivity in budding yeast. These data indicate that the yeast *Bcnt* ortholog is not essential for survival, but contributes to maintaining physiological homeostasis at the transcriptional level.

Whereas *Bcnt*-C is highly conserved among almost all eukaryotes, the N-terminal regions are less conserved. For example, the amino acids in *Drosophila Bcnt* (YETI) are ~50% identical to those of bovine *Bcnt/Cfdp1* in the C-terminal region, while the N-terminal region shows only ~22 % identity. Thus, although YETI is reported to bind to microtubule-based motor kinesin-I (Wisniewski, et al., 2003), a reevaluation is needed to confirm whether vertebrate *Bcnt* functions in intracellular trafficking because its interaction is mediated via its N-terminal region. One characteristic of the three *Bcnt*-related proteins is their different numbers of IR units: *Bcnt/Cfdp1*, *p97Bcnt/Cfdp2*, and *p97Bcnt2* have one, two, and three IR units, respectively. Sequences homologous to the 40 amino acid IR unit are found in zebra fish (*Danio rerio*) and nematodes, but not in yeast. These IR units comprise intrinsically disordered regions that might present scaffolds for protein-protein interaction as described later.

### 4.2 Intrinsic disorder of the three *Bcnt*-related proteins and cellular localization

The three *Bcnt*-related proteins move more slowly in sodium dodecyl sulfate acrylamide gel electrophoresis (SDS-PAGE) than expected, resulting in apparently higher molecular masses than those calculated. For example, bovine brain *Bcnt/Cfdp1* has 297 amino acids and a calculated molecular mass of 33.3 kDa, but appears around 45 kDa in SDS-PAGE (Fig. 6). The situation is exactly the same for both the *p97Bcnt/Cfdp2* and *p97Bcnt2* proteins, which have calculated molecular masses of 66.3 and 70.8 kDa, respectively (Iwashita et al., 2003, 2009). This might be caused by their physical properties in that the three *Bcnt*-related proteins are intrinsically disordered proteins (IDPs). It has been shown that many biologically active proteins lack a stable three-dimensional (3-D) structure; such proteins are referred to as IDPs (Dunker et al., 2008). IDPs are common to the three domains of life, and, especially in multicellular eukaryotic proteins, account for more than 70% of total proteins. They are involved in the regulation of various signalings through protein-protein

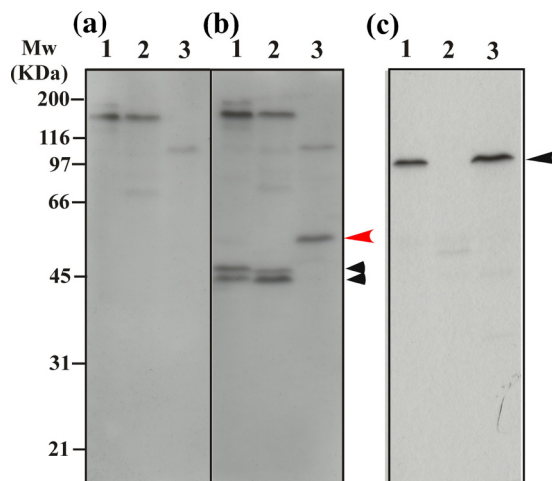


Fig. 6. Unique mobility of the Bcnt/Cfdp1 and p97Bcnt/Cfdp2 proteins in SDS-PAGE

Extracts of bovine brain (1), rat brain (2) and MDBK cells, a bovine kidney epithelial cell line (3) were separated in SDS polyacrylamide gels and subjected to immunoblotting with anti-Bcnt-C peptide antibody in the presence (a) or absence (b) of antigen peptide at a final concentration of 100  $\mu$ M, or with anti-p97Bcnt monoclonal antibodies (c). The two small black arrows indicate Bcnt/Cfdp1 with an apparent molecular mass of 45 kDa appearing as a doublet, probably due to phosphorylation (Iwashita et. al., 2003); the red large arrow indicates Bcnt/Cfdp1 with an apparent molecular mass of 53 kDa as described below

interactions that are frequently triggered by posttranslational modifications within the regions of intrinsic disorder (Dunker et al., 2008). IDPs may function as hub proteins via the formation of complexes with cellular proteins, which are then modulated by protein modifications such as phosphorylation, acetylation, ubiquitination, or degradation.

By computational prediction, Bcnt/Cfdp1, p97Bcnt/Cfdp2, and p97Bcnt2 are all suggested to comprise intrinsically disordered regions, except for the core of RTE domains that correspond to AP-endonuclease in the two paralogs (Fig. 7). This computational prediction is partially supported by an NMR study of the 3-D structure of the N-terminal 40 amino acid residues of the Bcnt-C region prepared in *Escherichia coli* using  $^{15}$ N-labeled amino acids. The spectrum revealed a lack of fixed tertiary structure (courtesy of Dr. T. Kohno). Furthermore, the Bcnt/Cfdp1 protein forms a tight protein complex with cellular proteins in bovine placenta even in the presence of a detergent, CHAPSO, when evaluated by gel filtration chromatography on Sephacryl S-300 HR followed by western blotting. Both Bcnt/Cfdp1 and p97Bcnt/Cfdp2 are phosphoproteins that are potentially phosphorylated on serine residues by casein kinase II *in vitro* (Iwashita et. al., 1999). Recently, the two phosphorylated serine residues in human BCNT,  $^{116}$ S in the N-terminal region and  $^{250}$ S in the C-terminal region, were identified by mass spectrometric analysis (Dephoure et al., 2008). This phosphorylation is cell cycle independent. It should be noted that these two phosphorylated serine residues reside in amino acid sequences WASF and WESF, respectively, which implies a unique motif for specific phosphorylation. Phosphorylation on these motifs might be expected to play a role in switching, such as switching the cation- $\pi$  mediated protein-ligand interaction (Zacharias & Dougherty, 2002). These characteristics suggest that the three Bcnt-related family members are hub-like molecules.



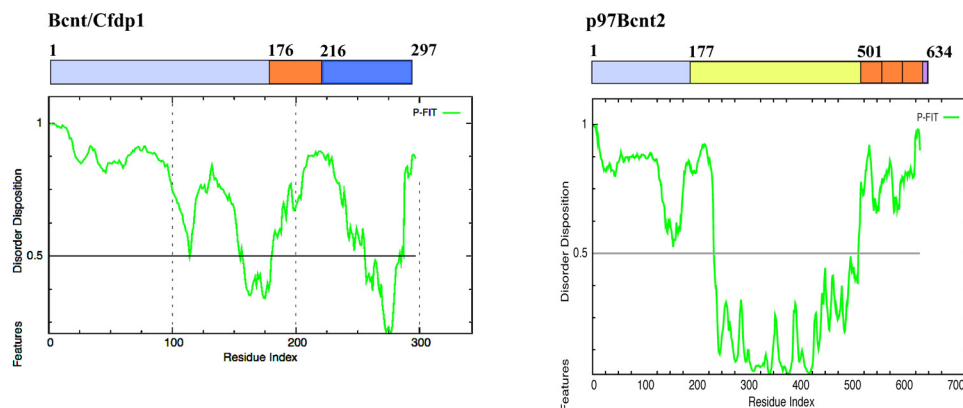


Fig. 7. Characteristics of intrinsic disorder of three Bcnt-related proteins

The Bcnt/Cfdp1, p97Bcnt/Cfdp2, and p97Bcnt2 proteins are predicted to comprise intrinsically disordered regions. Amino acid sequences of the three Bcnt-related proteins were subjected to analysis by a soft server of DisProt (Sickmeier et al., 2007), and individual profiles were obtained. The data for p97Bcnt/Cfdp2 are not shown, but are quite similar to those for p97Bcnt2. Vertical axes indicate the disorder probability of each amino acid residue, and the horizontal axes indicate the number of amino acid residues. Schematic domain structures of Bcnt/Cfdp1 and p97Bcnt2 are shown for each profile for comparison. Similar results were obtained using another program, Anchor (Mészáros et al., 2009)

We have found that the Bcnt/Cfdp1 protein from MDBK cells, a bovine kidney epithelial cell line (Madin & Darby, 1965; Iwashita et al., 1999), migrates at around 53 kDa in SDS-PAGE, significantly bigger than the rat or bovine brain proteins (Fig. 6). The same shift is observed in many other ruminant organs such as bovine placenta, testis and goat kidney, but not in all rat organs. Although we have yet not determined the cause, clarification of this anomaly could shed light on the role of Bcnt/Cfdp1, because the modification may be related to Bcnt/Cfdp1 function. Whereas the ~175 amino acid N-terminal regions of the three Bcnt-related proteins are acidic as a whole, they contain several arginine/lysine-rich elements, including a putative nuclear targeting motif of Arg-Lys-Arg-Lys (~61-64<sup>th</sup>). Therefore we examined the cellular distribution of the three Bcnt-related proteins in MDBK cells. The three were localized in both the cytosolic and nuclear fractions, and, in addition, both p97Bcnt/Cfdp2 and p97Bcnt2 were found significantly in the chromatin fractions (Fig. 8). These results suggest that Bcnt family members have the potential to function as shuttle molecules between the cytosol and nuclei. The nuclear localizations of p97Bcnt/Cfdp2 and p97Bcnt2 are consistent with their protein structure domains; the two paralogs include AP-endonuclease domains in the middle of the molecule as described in more detail below. On the other hand, either the 45 kDa (all rat organs and bovine brain) or 53 kDa (MDBK cells) Bcnt/Cfdp1 is scarcely found in the chromatin fraction, although chicken Bcnt/Cfdp1 has been reported as a centromere protein in a transformed cell line (Ohta et al., 2010).

#### 4.3 RTE domains of p97Bcnt/Cfdp2 and p97Bcnt2

AP-endonuclease is well known to function as an abasic endonuclease in the base excision repair pathway. It possesses multiple enzymic activities as a 3'-5' DNA exonuclease,



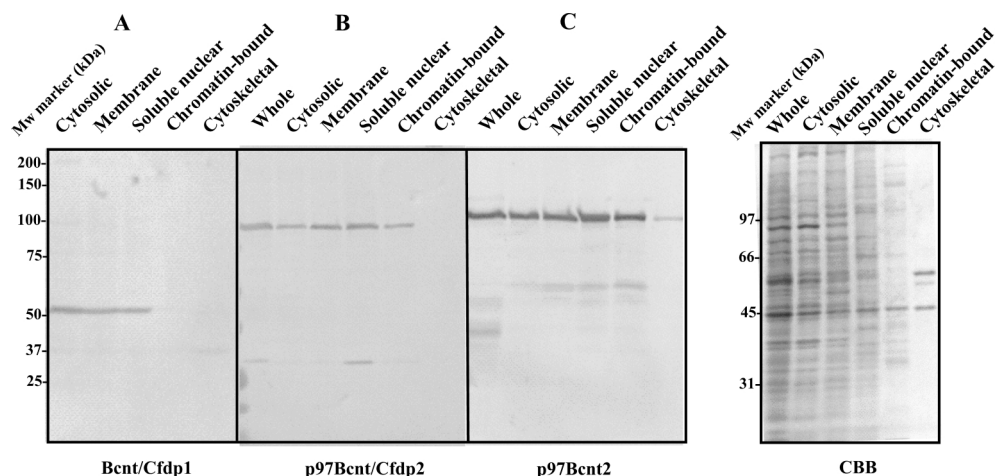


Fig. 8. Subcellular distribution of the three Bcnt-related proteins in MDBK cells

Subcellular fractionation of cultured MDBK cells was carried out successively using a Subcellular Protein Fractionation Kit from Pierce. Constant volume amounts of each fraction were assessed by immunoblotting. A: anti-Bcnt-C peptide antibody (Iwashita et al., 2003), B: anti-p97Bcnt monoclonal antibodies (Nobukuni et al., 1997) and C: anti-p97Bcnt2 peptide antibody (Iwashita et al., 2009). The right panel shows the Coomassie Brilliant Blue staining pattern. The subcellular fractions are identified at the top of the panels. The effectiveness of cellular fractionation was evaluated by immunoblotting using three antigens; anti-p120GAP (a marker for the cytosolic fraction, Kobayashi et al., 1993), anti-Topoisomerase II (a marker for the nuclear fraction, Iwashita et al., 1999), and anti-actin (a marker for the cytoskeleton). The data are consistent with previously reported results (not shown)

3'-phosphodiesterase, 3'-phosphatase, and RNase H (Barzilay et al., 1995). Many organisms possess two functional AP-endonucleases, which are thought to be important for cell viability. In contrast to non-vertebrate AP-endonuclease, vertebrate AP-endonuclease, which has an extra 6 kDa N-terminal region of intrinsic disorder, plays a role not only in repairing DNA damage, but also in regulating the redox state of various proteins that modulate transcription factors such as AP-1 (Fos/Jun), NF- $\kappa$ B, HIF-1, and p53 (Tell et al., 2009; Busso et al., 2010); thus it is termed APE/Ref-1 (AP-endonuclease/Redox effector factor 1). This is natural considering that DNA damage is one of the most vital stresses faced by living organisms. The extra N-terminal region of human AP-endonuclease (APE1) contains multiple arginine/lysine rich elements, and provides a scaffold for protein-protein interaction for DNA repair proteins such as Pol B and XRCC, and transcription factors including STAT3, YB-1, and nucleophosmin (NPM1) (Vascotto et al., 2009; Busso et al., 2010). Although we have not yet found evidence that p97Bcnt/Cfdp2 and p97Bcnt2 possess any of these activities, they have several characteristics common to mammalian AP-endonuclease with intrinsic disorder regions at both the N- and C-termini. Amino acid sequences in a part of the RTE domains are well conserved in all ruminants so far examined including *Lesser Malay chevrotrain* (Iwashita et al., 2009). The central 239-amino acid region of the RTE domain (termed the core RTE domain) corresponds exactly to Endonuclease/Exonuclease/Phosphatase family members (<http://www.ncbi.nlm.nih.gov/>

cdd?term=Pfam03372). The amino acid sequences of p97Bcnt/Cfdp2 and p97Bcnt2 were compared with those of three canonical AP-endonucleases: human APEX1, *Archaeoglobus* Af\_Exo, and *Neisseria* Nape (Fig. 9). Although the comparison revealed low overall identity (~20%) in the core RTE domains, eight amino acid residues involved in catalytic activity and at least 6 amino acids participating in substrate binding are conserved among the molecules. Furthermore, their 3-D structures could be remodeled with high accuracy, revealing the characteristics of Exo III or AP-endonuclease.

p97Bcnt	EYCIGT <b>W</b> <sup>*</sup> NVRSMPGKLDVVQEMERINIDILGISELKWGTG-----MGELNSDDHYIYY 295
p97Bcnt2	EYCIGT <b>W</b> NVRSMPGKLDVVQEMERINIDILGISELKWGTG-----MGELNSDDHYIYY 297
APEX1	TLKIC <b>S</b> W <b>N</b> VDGLRAWIKKKGLDWVKEEAPDILC <b>Q</b> ETKCSENKLP <b>A</b> ELQ <b>E</b> LPLSHQYWS 120
Nape	MLKII <b>S</b> ANVNGIRSAYKKGFY <b>E</b> YIAASGADIVCV <b>Q</b> ELKAQEADLSADMKNPHGMHGHWHC 60
Af_Exo	MLKIAT <b>F</b> NVNSIRSR-LHIVIPWLKENKPDILCM <b>Q</b> ETKVENRK----FPEADFHRIGYHV 55
p97Bcnt	CGQQSLRRNGV <b>A</b> LIVNKRVRNA <b>I</b> GCNLKNDRMISVR <b>F</b> QGGKPFNLTVIQV <b>Y</b> APT <b>P</b> Y-A <b>E</b> 354
p97Bcnt2	CGQQSLRRNGV <b>A</b> LIVNKRVRNA <b>I</b> GCNLKNDRMISVR <b>F</b> QGGKPFNLTVIQV <b>Y</b> APT <b>P</b> Y-A <b>E</b> 356
APEX1	APSDKEGYSVGLLSRQCPLKVSYGIGDEEHQ <b>E</b> GRVIVAEFDSFVLVTA <b>Y</b> VPNAG-RGL 179
Nape	AE--KRGS <b>G</b> VAVYSKRKPDNVQIGMGIEEFDR <b>E</b> GRFVRCD <b>F</b> GRLSV <b>I</b> SL <b>L</b> PSGS-SAE 117
Af_Exo	VFSGSKGRNGVAIASLEEPEDV <b>S</b> FGLDSEPKD-EDRLIRAKIAGIDVINT <b>V</b> PQGFKIDS 114
p97Bcnt	<b>G</b> E---VYRFYEDLQHLLE-ITPKIDVLF <b>I</b> GD <b>W</b> NAKVG <b>S</b> QEIP <b>G</b> ITG-RFLGMQNEAGR 409
p97Bcnt2	<b>P</b> E---VYRFYEDLQHLLE-ITPKIDVLF <b>I</b> GD <b>W</b> NAKVG <b>S</b> QEIP <b>R</b> ITG-KFGLGVQNEAGR 411
APEX1	VRLEYRQRWDEAFRKFLKGLASR-KPLVLC <b>G</b> DL <b>N</b> VAHEEIDLRNPKGNKKNAGT <b>P</b> QERQ 238
Nape	ERQQVKYRFLDAFYPMLEAMKNEGRD <b>I</b> VVCG <b>D</b> W <b>N</b> IAHQNIIDLKNWKG <b>N</b> QKNSGFLPEERE 177
Af_Exo	EKYQYKQLWLERLYHYLQKTVD <b>F</b> RSFAVWCG <b>D</b> M <b>N</b> VAPEPIDVHSPDKLNHVX <b>F</b> HEDARR 174
p97Bcnt	RLIEFC <b>H</b> HNRLV <b>I</b> NTL <b>F</b> QQ <b>P</b> SRRLY <b>T</b> WTSP-DGRYRD-----Q <b>I</b> DY <b>I</b> ICRQRWRSSVQS 463
p97Bcnt2	RLIEFC <b>Q</b> NRVL <b>I</b> ANTL <b>F</b> QQ <b>H</b> KRLY <b>T</b> WTSP-DGRYRD-----Q <b>I</b> DY <b>I</b> ICRQRWRSSVQS 465
APEX1	GFGELLQAVPLADSFRLHPNTPYAT <b>F</b> WTY-MMNARSKNVGWRL <b>D</b> YFLLSHSL <b>L</b> PALCD 297
Nape	WIGKVIHKLGTDMWRTLYPDVP-GYTWS <b>N</b> -RGQAYAKDVGW <b>R</b> I <b>D</b> YQ <b>M</b> VTPELA <b>A</b> AKAVS 235
Af_Exo	AYKKILELG-FVDVLRKIHPNER-IYTFYD <b>R</b> VKGAIERGLGW <b>R</b> G <b>A</b> ILAT <b>P</b> PLA <b>E</b> RCVD 232
p97Bcnt	AKTRPG---ADCG <b>S</b> <b>D</b> HKLLIAKF----- 483
p97Bcnt2	AKTRPG---ADCG <b>S</b> <b>D</b> HKLLIAKF----- 485
APEX1	SKIRS---KALGS <b>D</b> HCPITLYLAL--- 318
Nape	AHVYK---DEKFS <b>D</b> HAPLVVEYDYAAE 259
Af_Exo	CYADIKPR <b>L</b> AEKPS <b>D</b> HLLPLVAVFDV--- 257

Fig. 9. Highly conserved amino acid residues of the core RTE domains critical for AP-endonuclease activity

The amino acid sequences of the core RTE domains of p97Bcnt/Cfdp2 (241-483<sup>th</sup>) and p97Bcnt2 (243-485<sup>th</sup>), APEX1 (human APE, 61-318<sup>th</sup>), Nape (*Neisseria*, 1-259<sup>th</sup>, Carpenter et al., 2007) and Af\_Exo (*Archaeoglobus fulgidus*, 1-257<sup>th</sup>, Schmiedel et al., 2009) were aligned by the ClustalW2 program of EMBL-EBI. Residues critical for the catalytic activity of canonical AP-endonucleases are shown in red bold, and amino acid substitutions in the core RTE domains between p97Bcnt/Cfdp2 and p97Bcnt2 are indicated in blue bold

The 3-D structure of AP-endonuclease is evolutionarily well conserved and comprises two domains, each containing six-stranded  $\beta$  sheets decorated by helices on the concave site (Barzilay et al., 1995). The predicted 3-D structures of both RTE domains of p97Bcnt/Cfdp2 and p97Bcnt2 are quite similar to each other, and present possible DNA-binding sites between  $\alpha$ -helix domains and opposite both the N-terminal and C-terminal regions of the RTE domain. Next we superimposed the structure of p97Bcnt2 onto that of p97Bcnt/Cfdp2 to examine the structural relationship between the two domains (Fig. 10). Whereas the N-terminal ~60-amino acid region is variable between the two, there are only 12 amino acid differences in the 239-amino acid core RTE domain. It is noteworthy that 7 of the 12 different residues are located in the neighborhood of the predicted active sites. It is characteristic that the enzymatic properties of AP-endonuclease change significantly with subtle changes in the neighborhood of the active cavity. For example, a single amino acid substitution restores

Neisserial AP-endonuclease activity from the exonuclease (Carpenter et al., 2007), and a spontaneous substitution of Val to Gly in the C-terminal *Archaeoglobus* AP-endonuclease, which participates in forming an abasic DNA binding pocket, is accompanied by an increase in non-specific endonuclease activity (Schmiedel et al., 2009). This is probably because AP-endonuclease possesses multiple enzymatic activities as described above. Thus it could be expected that p97Bcnt/Cfdp2 and p97Bcnt2 would have different enzymatic properties, with each compensating for the function of the other.

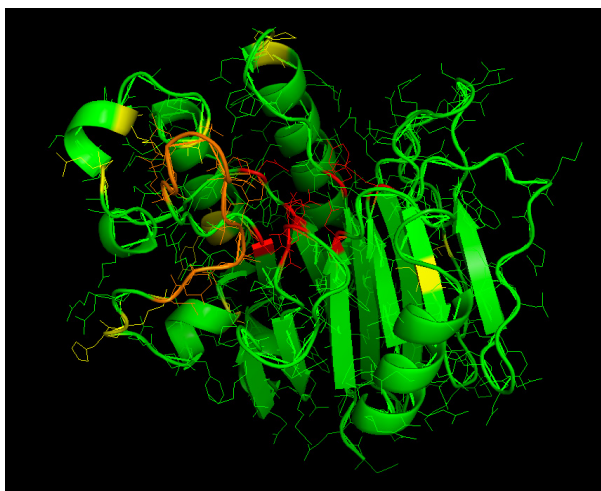


Fig. 10. 3-D comparison of the two core RTE domains of p97Bcnt/Cfdp2 and p97Bcnt2

3-D structures of the two core RTE domains of p97Bcnt/Cfdp2 and p97Bcnt2 were remodeled by the I-TASSER server (Roy et al., 2010), and p97Bcnt2 was superimposed on the p97Bcnt axis fixing the main alpha chain. RMSD (root mean square deviation) of 243 amino acid residues was 0.68 Å. From the data of the top templates of 2jc5A (Carpenter et al., 2007) and 2voaA (Schmiedel et al., 2009), the catalytic sites and DNA binding sites in red, or a loop involved in the targeting site in orange from a template of 2v0rA (Repanas et al., 2007) were identified. Twelve amino acid substitutions between the two domains are shown in yellow. These drawings were obtained using PyMOL. Analysis was carried out courtesy of Dr. M. Tanio, National Institutes of Natural Sciences, Okazaki

To explore further whether the nucleotide substitutions in p97Bcnt2 reflect natural selection in the two paralogs, we examined the  $d_N$  (non-synonymous substitution per site)/ $d_S$  (synonymous substitution per site) values for both the core RTE domain (243-483<sup>th</sup> amino acids) and the remaining regions (177-242<sup>th</sup> and 484-500<sup>th</sup> amino acids). The  $d_N/d_S$  values are 0.029/0.160 in the core region and 0.166/0.414 in the other regions. Although there are more non-synonymous and synonymous substitutions outside the core RTE domain of p97Bcnt2, the  $d_N/d_S$  values are  $< 1$ , suggesting no definite attribution to positive selection. On the other hand, the  $d_N/d_S$  value in the core RTE domain is much lower than that of the other regions, suggesting that selective constraints have been substantially strong in the core RTE domain (Iwashita et al., 2009). These data suggest that the recruited RTE domains in both p97Bcnt/Cfdp2 and p97Bcnt2 have played a crucial role in the duplicated novel genes.

## 5. Perspectives

Living organisms have evolved so as to acquire various anti-stress systems in response not only to exogenous stresses, but also to the intrinsic stresses faced by multicellular organisms for physiological homeostasis. DNA repair systems in all organisms, immune systems in vertebrates, and the placental systems in mammals are some of the most fruitfully acquired systems. Several lines of evidence have shown that the expression of various integrated retrotransposons is induced by environmental stimuli, such as ultraviolet light, heat shock, or macromolecule synthesis inhibitors (Liu et al., 1995; Morales et al., 2003; Häslér et al., 2007). Although the induction mechanism of expression has not been fully elucidated, these stresses may enhance promoter activity (Morales et al., 2003) or release the suppressive state of expression, resulting in the creation of new genetic materials. If the retrotransposon induction process is combined with tandem gene duplication, it is a much more efficient way to create a novel gene. Under such stressful conditions, novel genetic materials may play a role in adaptation to new environments.

Among the three Bcnt-related proteins in ruminants, the two paralogs p97Bcnt/Cfdp2 and p97Bcnt2 were generated by partial segmental duplication of the ancestral *Bcnt/Cfdp1* gene, followed by the insertion of an order-specific retrotransposon, resulting in the recruitment of the AP-endonuclease domain of the retrotransposon. Based on the 3-D remodeled structures of these recruited RTE domains and comparison of their protein sequences, they probably retain AP-endonuclease activity. Mammalian AP-endonuclease plays a role not only in direct DNA repair, but also in stimulating pathways for anti-stress activities. Although the latter activity depends mostly on the N-terminal 6-kDa region (Tell et al., 2009; Busso et al., 2010) which is dissimilar to those of p97Bcnt/Cfdp2 and p97Bcnt2, they contain intrinsically disordered regions other than the core RTE domains in both the N-terminal and C-terminal regions, including outside of the core RTE domains. These regions may serve as scaffolds for cellular protein-protein interactions and create novel functions as chimeric genes. The two paralogs distribute in both the cytosolic and nuclear fractions. The properties of intrinsically disordered proteins other than the AP-endonuclease domain and wide cellular distribution are similar to those of the APE1/Ref-1 molecule. Based on these considerations, we conclude that p97Bcnt/Cfdp2 and p97Bcnt2 have recruited the AP-endonuclease domain of a retrotransposon, which originally played an essential role in the integration of the retrotransposon into the genome, and that the two paralogs may have utilized AP-endonuclease activity to suppress cellular stress for survival. The cellular stress that might have induced the retrotransposition of Bov-B LINEs would increase the probability that the newly created genes would become fixed in a population. In addition, because these novel genes have chimeric origins, the original regulatory network of the ancestral *Bcnt/Cfdp1* gene may also have been modified to some extent. Therefore, we hypothesize that the two novel genes have become additional components of pre-existing regulatory networks for anti-stress activities. Although this hypothesis cannot explain why molecules containing the AP-endonuclease domain, such as p97Bcnt/Cfdp2 and p97Bcnt2, are so rare despite the advantage of being able to regulate cellular activity, it will be intriguing to examine the functions of the three Bcnt-related proteins based on this working hypothesis.

## 6. Conclusion

The *Bcnt/Cfdp1* gene comprises a unique gene family with three members in ruminants. The two paralogs, p97Bcnt/cfdp2 and p97Bcnt-2, were created in ancient ruminants by a partial

duplication of the ancestral *Bcnt/Cfdp1* gene followed by the insertion of an order-specific retrotransposon, Bov-B LINE. This type of combined process provides great potential to generate a novel gene because a novel function can be acquired under the guarantee of the original gene function. The ancestral *Bcnt/Cfdp1* protein contains a highly conserved C-terminus of 82-amino acids (*Bcnt-C*) that is not present in either *p97Bcnt/Cfdp2* or *p97Bcnt2*. *Bcnt-C* is found in all eukaryotes where it is known as the BCNT superfamily. A chicken *Bcnt/Cfdp1* is a centromere protein while the yeast ortholog is a component of the chromatin-remodeling complex, suggesting that the ancestral *Bcnt/Cfdp1* protein plays a role in the regulation of gene expression. The two paralogs, *p97Bcnt/Cfdp2* and *p97Bcnt2*, recruited an AP-endonuclease domain of the retrotransposon during their generation process as a ~325 amino acid region (RTE domain) in the middle of the molecule. The three *Bcnt*-related proteins distribute in both the cytosolic and nuclear fractions, and include intrinsically disordered regions other than the core of RTE domains of the two paralogs. The 3-D structures of the core RTE domains can be remodeled as canonical AP-endonucleases with identical catalytic amino acid residues. Although as yet there is no direct evidence for it, the two paralogs probably retain AP-endonuclease activity. Because AP-endonuclease/Redox effector factor 1 is one of the major regulators of cellular responses to various stresses, we propose that the recruited AP-endonuclease domains, which may have emerged in response to cellular stresses, may be utilized by the paralogs in cellular regulation. Therefore, the three *Bcnt*-related family members provide a good opportunity to examine dynamic changes in signaling networks that accompany novel genes.

## 7. Acknowledgements

We thank all our colleagues for their contributions to the study over the last 15 years, especially to Drs. K. Hashimoto and S. Hattori for their indispensable help in the early stages. We are grateful to Dr. T. Kohno for providing unpublished data, to Drs. H. Ohmori, M. Tanio, S.-Y. Song, K. Nakashima, S. Imajo-Ohmi, E. B. Kuettner, M.B. Gerstein, G. Tell, Y. Miyata, and Y. Ohno-Iwashita, and to The I-TASSER Server Team for useful discussion, and to Dr. D. Izumi for providing unpublished data. We are also grateful to Dr. Y. Nagai, the former president of Mitsubishi Kagaku Institute Life Sciences, for continuous encouragement, and to Dr. M. Dooley-Ohto for patient editing. During the preparation of this article, one of the authors in northern Japan suffered the disasters of a major earthquake and resulting tsunami, which led to the Fukushima nuclear plant accident. In this, we recognize the power of nature, and realize the importance of observing it carefully and describing it correctly.

## 8. References

- Barzilay, G., Walker, L.J., Robson, C.N., & Hickson, I.D. 1995. Site-directed mutagenesis of the human DNA repair enzyme HAP1: identification of residues important for AP endonuclease and RNase H activity. *Nucleic Acids Res.*, 23, pp. 1544-1550
- Brosius, J. 2005. Echoes from the past—are we still in an RNP world? *Cytogenet. Genome Res.*, 110, pp. 8-24
- Busso, C.S., Lake, M.W., & Izumi, T. 2010. Posttranslational modification of mammalian AP endonuclease (APE1). *Cell. Mol. Life Sci.*, 67, pp. 3609-3620

- Carpenter, E.P., Corbett, A., Thomson, H., Adacha, J., Jensen, K., Bergeron, J., Kasampalidis, I., Exley, R., Winterbotham, M., Tang, C., Baldwin, G.S., & Freemont, P. 2007. AP endonuclease paralogues with distinct activities in DNA repair and bacterial pathogenesis. *EMBO J.*, 26, pp. 1363-1372
- Dephoure, N., Zhou, C., Villén, J., Beausoleil, S.A., Bakalarski, C.E., Elledge, S.J., & Gygi, S.P. 2008. A quantitative atlas of mitotic phosphorylation. *Proc. Natl. Acad. Sci. USA*, 105, pp. 10762-10767
- Diekwisch, T.G., Marches, F., Williams, A., & Luan, X. 1999. Cloning, gene expression, and characterization of CP27, a novel gene in mouse embryogenesis. *Gene*, 235, pp. 19-30
- Dosztányi, Z., Mészáros, B., & Simon, I. 2009. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, 25, pp. 2745-2746
- Dunker, A.K., Oldfield, C.J., Meng, J., Romero, P., Yang, J.Y., Chen, J.W., Vacic, V., Obradovic, Z., & Uversky, V.N. 2008. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* 9, Suppl 2:S1.
- Fritz, G., & Kaina, B., 1999. Phosphorylation of the DNA repair protein APE/REF-1 by CKII affects redox regulation of AP-1. *Oncogene*, 18, pp. 1033-1040
- Gentles, A.J., Wakefield, M.J., Kohany, O., Gu, W., Batzer, M.A., Pollock, D.D., & Jurka, J. 2007. Evolutionary dynamics of transposable elements in the short-tailed opossum *Monodelphis domestica*. *Genome Res.*, 17, pp. 992-1004
- Hastings, P. J., Lupski, J. R., Rosenberg, S.M., & Ira, G. 2009. Mechanisms of change in gene copy number. *Nature Reviews. Genetics*, 10, pp. 551-564
- Iwashita, S., Nobukuni, T., Tanaka, S., Kobayashi, M., Iwanaga, T., Tamate, H.B., Masui, T., Takahashi, I., & Hashimoto, K. 1999. Partial nuclear localization of a bovine phosphoprotein, BCNT, that includes a region derived from a LINE repetitive sequence in Ruminantia. *Biochim. Biophys. Acta*, 1427, pp. 408-416
- Iwashita, S., Osada, N., Itoh, T., Sezaki, M., Ohshima, K., Hashimoto, E., Kitagawa-Arita, Y., Takahashi, I., Masui, T., Hashimoto, K., & Makalowski, W. 2003. A transposable element-mediated gene divergence that directly produces a novel type bovine Bcnt protein including the endonuclease domain of RTE-1. *Mol. Biol. Evol.*, 20, pp. 1556-1563
- Iwashita, S., Ueno, S., Nakashima, K., Song, S.-Y., Ohshima, K., Tanaka, K., Endo, H., Kimura, J., Kurohmaru, M., Fukuta, K., David, L., & Osada, N. 2006. A tandem gene duplication followed by recruitment of a retrotransposon created the paralogous bucentaur gene (*bcntp<sup>97</sup>*) in the ancestral ruminant. *Mol. Biol. Evol.*, 23, pp. 798-806
- Iwashita, S., & Song, S.-Y. 2008. RasGAPs: a crucial regulator of extracellular stimuli for homeostasis of cellular functions. *Mol. BioSyst.*, 4, pp. 213-222
- Iwashita, S., Nakashima, K., Sasaki, M., Osada, N., & Song, S.-Y. 2009. Multiple duplication of the bucentaur gene family, which recruits the APE-like domain of retrotransposon: Identification of a novel homolog and distinct cellular expression. *Gene*, 435, pp. 88-95
- Jobse, C., Buntjer, J.B., Haagsma, N., Breukelman, H.J., Beintema, J.J., & Lenstra, J.A. 1995. Evolution and recombination of bovine DNA repeats. *J. Mol. Evol.*, 41, pp. 277-283
- Kaessmann, H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20, pp. 1313-1326

- Kobayashi, M., Hashimoto, N., Hoshino, M., Hattori, S., & Iwashita, S. 1993. Differential contribution of Mr 120 kDa rasGTPase-activating protein and neurofibromatosis type 1 gene product during the transition from growth phase to arrested state in human fibroblasts accompanied by a unique rasGTPase-activating activity. *FEBS Lett.* 327, pp. 177-182
- Liu, W.M., Chu, W.M., Choudary, P.V., & Schmid, C.W. 1995. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res.*, 23, pp. 1758-1765
- Liu, G.E., Ventura, M., Cellamare, A., Chen, L., Cheng, Z., Zhu, B., Li, C., Song, J., & Eichler, E.E. 2009. Analysis of recent segmental duplications in the bovine genome. *BMC Genomics.*, 10, pp.571
- Makalowski, W. 2003. Not junk after all. *Science*, 300, pp. 1246-1247
- Malik, H.S., & Eickbush, T.H. 1998. The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. *Mol. Biol. Evol.* 15, pp. 1123-1134
- Madin, S.H., & Darby, N.B.Jr. 1958. Established kidney cell lines of normal adult bovine and ovine origin. *Proc. Soc. Exp. Biol. Med.* 98, pp. 574-576
- Morales, J.F., Snow, E.T., & Murnane, J.P. 2003. Environmental factors affecting transcription of the human L1 retrotransposon. II. Stressors. *Mutagenesis*, 18, pp. 151- 158
- Nobukuni, T., Kobayashi, M., Omori, A., Ichinose, S., Iwanaga, T., Takahashi, I., Hashimoto, K., Hattori, S., Kaibuchi, K., Miyata, Y., Masui, T., & Iwashita, S. 1997. An Alu-linked repetitive sequence corresponding to 280 amino acids is expressed in a novel bovine protein, but not in its human homologue. *J. Biol. Chem.*, 272, pp. 2801-2807
- Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, New York
- Ohta, S., Bukowski-Wills, J.-C., Sanchez-Pulido, L., Alves, de Lima Alves, F., Wood, L., Chen, Z.A. Platani, M., Fischer, L., Hudson, D. F., Ponting, C.P., Fukagawa, T., Earnshaw, W. C., & Rappsilber, J. 2010. The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics. *Cell*, 142, pp. 810-821
- Rawn, S.M., & Cross, J.C. 2008. The evolution, regulation, and function of placenta-specific genes. *Annu. Rev. Cell Dev. Biol.*, 24, pp. 159-181
- Repanas, K., Zingler, N., Layer, L.E., Schumann, G.G., Perrakis, A., & Weichenrieder, O. 2007. Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease. *Nucleic Acids Res.*, 35, pp. 4914-4926
- Roy, A., Kucukural, A., & Zhang, Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.*, 5, pp. 725-738
- Schmiedel, R., Kuettner, E.B., Keim, A., Sträter, N., & Greiner-Stöffe, T. 2009. Structure and function of the abasic site specificity pocket of an AP endonuclease from *Archaeoglobus fulgidus*. *DNA Repair (Amst)*, 8, pp. 219-231
- Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B, Tompa, P., Chen, J., Uversky, V.N., Obradovic, Z., & Dunker, A.K. 2007. DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.*, 35, D786-793
- Sorek, R. 2007. The birth of new exons: Mechanisms and evolutionary consequences. *RNA*, 13, pp. 1603-1608

- Szemraj, J., Plucienniczak, G., Jaworski, J., & Plucienniczak, A. 1995. Bovine Alu-like sequences mediate transposition of a new site-specific retroelement. *Gene*, 152, pp. 261-264
- Takahashi, I., Nobukuni, T., Ohmori, H., Kobayashi, M., Tanaka, S., Ohshima, K., Okada, N., Masui, T., Hashimoto, K., & Iwashita, S. 1998. Existence of a bovine LINE repetitive insert that appears in the cDNA of bovine protein BCNT in ruminant, but not in human, genomes. *Gene*, 211, pp. 387-394
- Tell, G., Quadrifoglio, F., Tiribelli, C., & Kelley, M.R. 2009. The many functions of APE1/Ref-1: not only a DNA repair enzyme. *Antioxid. Redox Signal.*, 11, pp. 601-620
- The Bovine Genome Sequencing and Analysis Consortium. 2009. The Genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, 324, pp. 522-528
- Vascotto, C., Cesaratto, L., Zeef, L.A., Deganuto, M., D'Ambrosio, C., Scaloni, A., Romanello, M., Damante, G., Tagliatella, G., Delneri, D., Kelley, M.R., Mitra, S., Quadrifoglio, F., & Tell, G. 2009. Genome-wide analysis and proteomic studies reveal APE1/Ref-1 multifunctional role in mammalian cells. *Proteomics*, 9, pp. 1058-1074
- Wisniewski, T.P., Tanzi, C.L., & Gindhart, J.G. 2003. The *Drosophila* kinesin-I associated protein YETI binds both kinesin subunits. *Biol. Cell*, 95, pp. 595-602
- Wu, W.-H., Wu, C.-H., Ladurner, A., Mizuguchi, G., Wei, D., Xiao, H., Luk, E., Ranjan, A., & Wu, C. 2009. N-terminus of Swr1 binds to histone H2AZ and provides a platform for subunit assembly in the chromatin remodeling complex. *J. Biol. Chem.*, 284, pp. 6200-6207
- Zacharias, N., & Dougherty, D.A. 2002. Cation- $\pi$  interactions in ligand recognition and catalysis. *Trends Pharmacol. Sci.*, 23, pp. 281-287
- Zhang, J. 2003. Evolution by gene duplication: an update. *Trends Ecol. Evol.*, 18, pp. 292-298
- Zhang, J., Dean, A.M., Brunet, F., & Long, M. 2004. Evolving protein functional diversity in new genes of *Drosophila*. *Proc. Natl. Acad. Sci. USA*, 101, pp. 16246-16250
- Zingler, N., Weichenrieder, O., & Schumann, G.G. 2005. APE-type non-LTR retrotransposons: determinants involved in target site recognition. *Cytogenet. Genome Res.*, 110, pp. 250-268
- Zupunski, V., Gubensek, F., & Kordis, D. 2001. Evolutionary dynamics and evolutionary history in the RTE clade of non-LTR retrotransposons. *Mol. Biol. Evol.*, 18, pp. 1849-1863